Using NLP Methods for the Analysis of Rituals

Nils Reiter♣, Oliver Hellwig♠, Anand Mishra♠, Anette Frank♣, Jens Burkhardt♣

♣ Department of Computational Linguistics, ♠ South Asia Institute
Heidelberg University, Germany
www.cl.uni-heidelberg.de, www.sai.uni-heidelberg.de
{reiter,frank,burkhardt}@cl.uni-heidelberg.de, hellwig7@gmx.de, anand.mishra@urz.uni-heidelberg.de

Abstract

This paper gives an overview of an interdisciplinary research project that is concerned with the application of computational linguistics methods to the analysis of the structure and variance of rituals, as investigated in ritual science. We present motivation and prospects of a computational approach to ritual research, and explain the choice of specific analysis techniques. We discuss design decisions for data collection and processing and present the general NLP architecture. For the analysis of ritual descriptions, we apply the frame semantics paradigm with newly invented frames where appropriate. Using scientific ritual research literature, we experimented with several techniques of automatic extraction of domain terms for the domain of rituals. As ritual research is a highly interdiciplinary endavour, a vocabulary common to all sub-areas of ritual research can is hard to specify and highly controversial. The domain terms extracted from ritual research literature are used as a basis for a common vocabulary and thus help the creation of ritual specific frames. We applied the tf*idf, χ^2 and PageRank algorithm to our ritual research literature corpus and two non-domain corpora: The British National Corpus and the British Academic Written English corpus. All corpora have been part of speech tagged and lemmatized. The domain terms have been evaluated by two ritual experts independently. Interestingly, the results of the algorithms were different for different parts of speech. This finding is in line with the fact that the inter-annotator agreement also differs between parts of speech.

1. Introduction

The structure and dynamics of rituals within and across different cultures and eras is the focus of a large interdisciplinary collaborative research center including 21 scientific fields ranging from Indology to Musicology.[1] The project presented in this paper complements traditional research methods prevalent in the humanities with computational linguistics analysis methods. In particular, we aim at employing data-driven approaches to detect regularities and variations of rituals, based on semi-automatic semantic annotation of ritual descriptions.

Section 2 will present motivations for a corpus-based computational linguistics approach to ritual structure research and the project research plan. Section 3 illustrates our approach for semantic annotation and structural analysis of ritual descriptions. In Section 4 we report on first attempts to automatically acquire ritual-specific terminology. Section 5 concludes.

2. Computational Linguistics for Ritual Structure Research

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

Our interdisciplinary project 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 descriptions of complex event sequences, involving designated participants, objects, places and times, usually encoded in natural language descriptions. However, the knowledge of patterns in ritual event sequences is often highly private among researchers devoted to particular cultures or scientific fields. Our project attempts to make these patterns overt, through computational linguistic analysis of the textual ritual descriptions.

Computational Linguistics has developed semantic lexica and processing tools for the formal analysis of events and their predicate-argument structure, in terms of semantic roles. Based on such structured and normalized semantic representations of event sequences, we can identify recurrent patterns and variations across rituals by quantitative analysis. Frame Semantics (Fillmore et al., 2003), with its concept of scenario frames connected by frame relations and role inheritance, offers a powerful framework for the analysis of complex event sequences in ritual descriptions. Through annotation of word senses we can observe and analyze variations in the selectional characteristics of specific events and their roles across rituals. Finally, the creation of structured and normalized semantic representations for ritual descriptions will allow 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.

For Computational Linguistics research, the project allows for the detailed investigations of techniques to compute event chains as representations of complex action sequences. Ritual descriptions typically consist of complex and often recurrent event sequences and use a restricted domain vocabulary and a closely circumscribed inventory of events and participants. This somewhat controlled domain gives us the opportunity for a detailed study of several phenomena at the interface between syntax and semantics (e.g. the computation of selectional preferences, the annotation of ritual-specific frames, and the modeling of event

[1] “Ritual Dynamics”: http://www.ritualdynamik.de

sequences in ontologies).

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.

2.1.1. Corpus creation and annotation

In the first stage, a comprehensive corpus of linguistically and semantically annotated rituals from different cultures is being created from natural language descriptions of rituals that are procured by experts. The semantic annotation follows the frame semantics paradigm (Fillmore et al., 2003) and comprises both general linguistic and ritual-specific annotation levels.

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.

2.1.2. 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 experts, we will provide tools and explore methods for empirical, quantitative analysis of rituals, based on abstract semantic representations of rituals.

2.2. Related Work

2.2.1. Event Chains

Central to the structure of rituals are sequences of events and participants involved in these events. Thus, 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 Burchardt et al. (2006). 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 co-reference or syntactic association. Recently, a statistical approach has been proposed by Chambers and Jurafsky (2009) for unsupervised detection of event chains, using co-occurrence of a single discourse entity as argument of different verbs as well as co-reference information as criteria for extracting event chains. A related shared task on "linking roles in discourse" (Ruppenhofer et al., 2009) is being organized as part of SemEval 2010.

2.2.2. Term Extraction

Term extraction for specific domains has been a field of study for quite some time. Usually, two or more corpora are compared: Terms that appear significantly more often in a domain corpus compared to a domain "neutral" corpus are considered as domain terms. Frank et al. (1999) and Buitelaar and Sacaleanu (2001) use tf$*$idf to measure term relevance; Agirre et al. (2001) propose the use of the χ^2-test. Reiter and Buitelaar (2008) successfully used χ^2 to detect medical domain terms.

An approach on domain term extraction that does not rely on contrasting corpora has been proposed by Yang et al. (2009). They transform the corpus into a graph containing nodes for candidate terms and edges between terms that appear in the same sentence. The PageRank algorithm is then used to weight the candidates, increasing the weight of term candidates co-occurring with term candidates that have a high weight in the same sentence.

3. Building an Annotated Corpus of Ritual Descriptions

As a basis for analysis, we use textual descriptions of rituals, which are supposed to include all relevant aspects of rituals.

3.1. Collection

We collect ritual descriptions from different sources. While some of them are found in lore (prescriptive sources), others were recorded in ethnological field studies (descriptive sources). This collection process has been started with rituals from Hinduism and Islam but we plan to adapt it for rituals from Ancient Egypt and the Middle Ages in central Europe.

3.1.1. Translations and Encodings

Ritual descriptions often contain specific lexemes that do not have a direct translation in English (or other languages):

(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 important to sweep with *kuśa* and not any other grass. As there is no English translation (the term "grass" refers to a more general concept), the translation is annotated with the original term. As the original name often contains non-Latin characters, the ritual descriptions are encoded in Unicode. For automatic processing, the original terms are eliminated, and later (re-)inserted in the semantic representation.

3.1.2. Fixed expressions

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

(2) While saying the mantra *oṃ sumitriyā na āpa [...]* he sweeps the head with the purifiers.

There is no common way to refer to these fixed expressions. Sometimes, prayers or chants have a title or name; sometimes, first words or the refrain are given and the practitioner (or expert) can infer what is meant.

As most fixed expressions cannot be directly translated, we adopt them as unanalyzed expressions in a foreign language. We ask the ritual experts to mark them as such, so that for processing we can replace them with indexed place holders and re-insert them into the semantic representation.

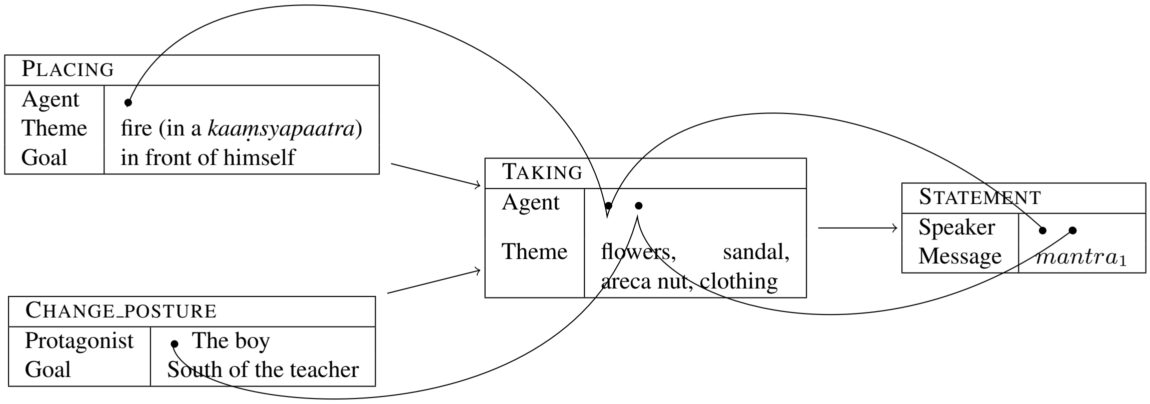

Figure 1: A schematic representation of a common subsequence in two different rituals

```
(S
  (NP (PRP He))
  (VP (VBZ sweeps)
    (NP
      (NP (DT the) (NN place))
      (PP (IN for)
        (NP
          (NP (DT the)
              (JJ sacrificial)
              (NN fire))
          (PP (IN with)
            (NP (NN grass)))))))
  (. .))
```

Figure 2: Parse tree for (1): "with grass" is attached to the NP "the sacrificial fire". This analysis clearly does not conform with the interpretation of the sentence.

3.1.3. Meta data

Various types of meta data need to be collected for different cultures. For instance, it is important to distinguish descriptive and prescriptive ritual sources. We store this information in a growing set of meta data with each ritual description, using standardized sets of meta data tags if applicable.

3.2. Linguistic Preprocessing

The ritual descriptions are preprocessed with standard NLP tools. We use UIMA[2] as a pipeline framework which currently consists of the Stanford POS-tagger (Toutanova et al., 2003), Named Entity Recognizer (Finkel et al., 2005) and Parser (Klein and Manning, 2003). The fact that none of the above tools are trained on ritual descriptions or comparable texts causes processing problems. At the time of writing, we cannot yet provide a detailed quantitative analysis of how these tools perform. However, we can identify PP-attachment as a major recurrent problem:

3.2.1. PP-Attachment

PPs are quite common in the data, as becomes apparent in Example (1). The Stanford parser shows a tendency to attach PPs to closer phrases rather than building large structures (see Figure 2).

In light of the frequent parse errors produced by the Stanford Parser, we currently experiment with using chunks produced by the OpenNLP Chunker as basis for the annotation.

3.3. Annotation

The annotation will be performed manually on syntactically parsed structures using the SALTO tool[3]. We will annotate two layers of semantic information.

3.3.1. General frame annotation

This annotation models the literal actions occurring while a ritual is performed. Ritual descriptions will be annotated with FrameNet frames, describing prototypical events together with their participants (denoted as frame elements). For most of the basic actions in rituals, an appropriate frame is already defined in FrameNet (80.8% of the verbs in the ritual descriptions collected so far [14,081 tokens] are known lexical units). For actions that are not represented in FrameNet, we will carefully add new frames and integrate them into the FrameNet hierarchy (e.g., frame relations, semantic types).

3.3.2. Ritual specific annotation

In addition to the literal actions they denote, most events have a ritual-specific meaning or intention (e.g., sweeping a place in order to purify it). We will add a second layer with ritual-specific frames from a newly created frame inventory. The ritual-specific frames will be linked to the general frames via the annotated lexical units and related to each other by frame relations where appropriate. This network of ritual frames will form the basis of a ritual ontology. A framework for the integration of an annotated corpus with an ontology has been presented by Burchardt et al. (2008).

3.4. Detecting Ritual Structure

As proof of concept for the types of analyses we can offer to ritual scientists, we constructed representations for a number of close variations of rituals. Figure 1 shows a partial semantic representation of two such rituals. We extracted the event sequences, one starting with PLACING, one with CHANGE_POSTURE. The sequences share the frames TAKING and STATEMENT. The co-reference chains are denoted

[2] http://incubator.apache.org/uima/

[3] http://www.coli.uni-saarland.de/projects/salsa/page.php?id=software

by curved lines. This is one way in which we plan to extract and visualize common subsequences in rituals.

4. Domain Term Extraction

As the ritual descriptions we collect come from different cultures, epochs and regions, the providing researchers come from various disciplines and speak different "scholarly languages". To support normalization of the used vocabulary, we collected a corpus of scientific literature from the various disciplines. From this we extract relevant terms for rituals *in general*, using three different approaches.

4.1. Corpora

Some of our approaches employ contrasting, non-domain corpora in order to identify domain terms. We use two different non-domain corpora, one general corpus (BNC) and one for scientific language with mixed subjects (BAWE).

4.1.1. BNC

The British National Corpus (BNC, 2007) consists of 100 million tokens from various domains and sources. We use both the written and spoken part of the BNC.

4.1.2. BAWE

The British Academic Written English corpus[4] contains 2761 documents written by students from various disciplines and levels of study (starting with undergraduate students) that were somewhat cleaned. In total, the corpus contains 6.3 million tokens. The corpus is already sentence split, but we applied automatic (heuristic) tokenizing, part-of-speech-tagging[5] and lemmatization (Toutanova et al., 2003).

4.2. Approaches

In the following, a term always includes its part of speech. Thus, the noun "worship" is a different term (and term candidate) than the verb "worship". As candidates, we used all nouns, verbs and adjectives occurring in the ritual literature corpus.

4.2.1. TF*IDF

The TF$*$IDF measure for termhood has been studied extensively in information retrieval. Let $\text{freq}_{t,d}$ be the frequency of term t in document d, df_t be the number of documents in which term t appears, and D the number of documents. The TF*IDF score of a term t in a document d is then calculated as shown in (3) and (4).

$$\text{tf}_{t,d} = \frac{\text{freq}_{t,d}}{\max_{t'} \text{freq}_{t',d}} \quad (3)$$

$$\text{tfidf}_{t,d} = \text{tf}_{t,d} * \log(\frac{D}{\text{df}_t}) \quad (4)$$

Dataset		Agreement			
		Positive		Negative	
		Full agr.	Partial agr.	Full agr.	Partial agr.
TF*IDF	All	13.33 %	23.33 %	16.67 %	56.67 %
	Nouns	10 %	20 %	0 %	70 %
	Verbs	20 %	30 %	0 %	40 %
	Adj.	10 %	20 %	50 %	60 %
χ^2	All	40 %	56.67 %	10 %	30 %
	Nouns	20 %	30 %	20 %	60 %
	Verbs	40 %	60 %	10 %	30 %
	Adj.	60 %	80 %	0 %	0 %
PageRank	All	10 %	20 %	13.33 %	60 %
	Nouns	20 %	30 %	10 %	50 %
	Verbs	0 %	0 %	10 %	60 %
	Adj.	10 %	30 %	20 %	70 %

Table 1: Ratio of domain terms in different data sets

As we are not aiming at identifying the most relevant document for a given term (as is the standard use case for TF*IDF), we need to slightly change the view on documents. Each corpus is combined in one document, so that we have three documents. Document frequency (df_t) is the number of corpora in which the term t appears.

4.2.2. Chi^2

For the χ^2 measure, the domain corpus was set in contrast to the BNC and BAWE corpora. We calculated χ^2 as described in Manning and Schütze (1999), but based on lemma- and POS-information and the summed frequencies over both non-domain corpora. The raw χ^2 values are logarithmized for normalization and scaled to the interval $[0, 1]$.

4.2.3. PageRank

The third approach on domain term extraction follows Yang et al. (2009). For each candidate term, a node in a graph is created. If two (candidate) terms co-occur in the same sentence, an edge between the corresponding nodes is added to the graph. The relevance of individual term nodes is then calculated using the PageRank (Brin and Page, 1998) algorithm. The PageRank algorithm gives higher weights to nodes that are connected to other nodes with a high weight – The relevance score of a term increases if it co-occurs with a term that has a high relevance score.

4.3. Evaluation

As there is no gold standard for the domain of ritual science, we asked two ritual experts to annotate the terms extracted by our approaches. From each approach, we selected the 10 best ranked noun, verb and adjective terms, so that in total 90 terms have been annotated.

The terms have been classified into three classes (yes, maybe, no) by two ritual experts independently. The overall kappa for this annotation is $\kappa = 0.35$, with differences between part-of-speech categories. The highest agreement between the annotators was achieved on adjectives ($\kappa =$

[4] The British Academic Written English (BAWE) corpus was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

[5] `http://opennlp.sf.net`

0.49). Interestingly, 36% of the agreed adjective terms are rejections. With a kappa value of 0.35, the agreement on verbs seems to be somewhat fair. The agreement was very low on nouns, with $\kappa = 0.22$. Because of this low agreement, we will also look at the data produced by the individual annotators A1 and A2 in the following discussion. In general, A1 seems to be much more liberal, annotating "maybe" in 58% of all cases (A2: 19%). The majority class for A2 is "no" (51%, A1: 14.4%). We also found inconsistencies within the data annotated by A1 (*granth*[6] is annotated as "maybe" in one approach and "no" in another).

Table 1 shows the results for the different term lists. Partial agreement is achieved when one annotator annotated yes (or no) and the other one maybe. In full agreement, both annotators annotated yes (or no).

From the terms extracted with the **TF*IDF** approach, 13.3% are considered a domain term by both annotators. As mentioned above, the annotator's judgement differs with respect to part of speech of the term. Among the nouns, only 10% of the terms are judged as domain terms by both annotators (verbs: 20%, adjectives 10%). If one annotator is allowed to be unsure, i.e., annotating the term as "maybe" (partial agreement), the number of terms increases to 23.3% (with similar increases for each part of speech). For wrongly extracted terms, the agreement between the annotators is much smaller. Among the adjectives, the annotators agreed that 50% of the terms are false positives, while there was no extracted noun or verb that was rejected by both annotators.

According to annotator A1, all the nouns and verbs extracted by the TF*IDF approach are certainly or maybe domain terms. Several very generic adjectives (*earlier*, ...) are annotated as non-domain in agreement with A2. A2 rejected most of the nouns, verbs and adjectives. Only 10% of the nouns, 20% of the verbs and 30% of the adjectives are annotated positively.

The overall numbers are better for the **Chi**^{2} approach. For 40% of the terms, both annotators fully agree that it is a domain term. 56.6% of the extracted terms are judged as domain relevant by at least one annotator (while the other specified "maybe"). Again, for different parts of speech, we get different results. While the results of the TF*IDF approach do not show a clear tendency, the χ^2 approach performs clearly better for adjectives than for verbs and nouns. If partial agreement is allowed, 80% of the adjectives are considered a domain term, while only 30% of the nouns and 60% of the verbs are judged as domain terms. Looking at the data, we find that the 20% of the nouns that are rejected by both annotators can both be traced to errors from the part of speech tagger: Both *www* and *varanasi*[7] are not verbs.

None of the annotators rejected any one of the adjectives. A1 annotated 60% of them as domain-term, A2 80%. Again, nouns seem to be much more problematic. Both rejected most of the nouns (A1: 30% yes; A2: 20% yes). For verbs, about half of the terms are accepted (A1: 40%; A2: 60%).

The **PageRank** approach scores much lower than the others for all three parts of speech. In fact, not a single verb extracted with this algorithm was judged as a domain term, even if partial agreement is allowed. Looking at the list of verbs, this is not surprising as most of them are very general (*to do*, *to be*, *to have*, ...). The same observation can be made for the other parts of speech. The extracted terms are clearly often used (and useful) in the domain, but do not separate the ritual research domain from others.

The annotations for the terms extracted using PageRank support what we mentioned above: Annotator A2 is much more strict. While A1 rejects only 13% of the terms, A2 rejects 60% across all word classes.

5. Conclusions

We presented motivations and the research plan of an interdisciplinary project that can offer insights for ritual science, but also for CL. We discussed problems we face when dealing with data from humanities, especially in the domain of rituals. We presented the work chain for corpus creation, annotation and exploitation for the structural analysis of rituals. We reported on preliminary results for the extraction domain terminology, which will provide a base vocabulary for ritual-specific semantic annotation. The next steps will be the semantic annotation of ritual descriptions on a larger scale and the deployment of analysis techniques to identify structural elements and variability in rituals.

Acknowledgements

We would like to thank our annotators Karin Polit and Christoph Zotter and our research assistants Irina Gossmann, Julio Rodrigues and Britta Zeller for their support. The project is part of the collaborative research center on ritual dynamics (SFB 619), funded by the German Research Foundation (DFG).

6. References

Eneko Agirre, Olatz Ansa, David Martinez, and Eduard Hovy. 2001. Enriching wordnet concepts with topic signatures. In *Proceedings of WordNet and Other Lexical Resources Workshop*.

BNC. 2007. The british national corpus, version 3 (bnc xml edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Paul Buitelaar and Bogdan Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In Dan Moldovan, Sanda Harabagiu, Wim Peters, Louise Guthrie, and Yorick Wilks, editors, *Proceedings of the Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations at NAACL*. NAACL, June.

Aljoscha Burchardt, Anette Frank, and Manfred Pinkal. 2006. Building Text Meaning Representations from Contextually Related Frames. In *Proceedings of IWCS*.

Aljoscha Burchardt, Sebastian Pado, Dennis Spohr, Anette Frank, and Ulrich Heid. 2008. Constructing integrated

[6] Adi *Granth* is the holy scripture of the Sikhs.

[7] Varanasi is a city in Northern India and thus a proper name.

corpus and lexicon models for multi-layer annotations in owl dl. *Linguistic Issues in Language Technology*, 1(1):1–33.

Nathanael Chambers and Dan Jurafsky. 2009. 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*, pages 602–610, Suntec, Singapore, August. Association for Computational Linguistics.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*.

Dan Klein and Christopher Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Nils Reiter and Paul Buitelaar. 2008. Lexical Enrichment of a Human Anatomy Ontology using WordNet. In *Proceedings of the Global WordNet Conference*.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2009. 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)*, pages 106–111, Boulder, Colorado, June. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL*, pages 252–259.

Yuhang Yang, Tiejun Zhao, Qin Lu, Dequan Zheng, and Hao Yu. 2009. Chinese term extraction using different types of relevance. 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*, pages 213–216.