

TOWARDS AUTOMATIC ANNOTATION OF TEXT TYPE STRUCTURE: EXPERIMENTS USING AN XML-ANNOTATED CORPUS AND AUTOMATIC TEXT CLASSIFICATION METHODS

Hagen Langer*, Harald Lungen†, Petra Saskia Bayerl†

* Justus-Liebig-Universität, Gießen
and Universität Osnabrück, Germany
hagen.langer@web.de

† Justus-Liebig-Universität, Gießen
Otto-Behaghel-Strasse 10D
35390 Giessen, Germany
{harald.luengen, petra.s.bayerl}@uni-giessen.de

Abstract

Scientific articles exhibit a fairly conventionalized structure in terms of topic types such as *background*, *researchTopic*, *method* and their ordering and rhetorical interrelations. This paper describes an effort to make such structures explicit by providing a corpus of German linguistic articles with XML markup according to a text type schema defining 21 topic type categories. The corpus is further augmented with XML annotations on a grammatical level and a logical structure level. The efficiency of an automatic annotation of text type structure is explored in experiments that apply general, domain-independent automatic text classification methods to text segments and employ features from the raw text level and the corpus annotations on the grammatical level. The results indicate that some of our topic types are successfully learnable.

1. INTRODUCTION

Nowadays the majority of scientific articles is published digitally in electronic libraries, on CDROMs/DVDs, and notably in the W3, e.g. on conference or journal sites, in online archives, or on researchers' home pages. The vast amount of information that is available via the new media at practically every point in time has afforded new techniques for goal-oriented search and retrieval, amongst other things, of scientific articles. Besides simple search via character strings in texts, many techniques require phases of pre-processing articles, e.g. by automatic classification, summarization or generation of metadata. One such technique is the CiteSeer approach of providing access to scientific articles on the W3 via automatically generated citation networks (Giles, Bollacker, & Lawrence, 1998). Another one is the categorization of whole articles into thematic categories such as scientific disciplines and sub-disciplines either automatically (e.g. Sebastiani, 2001) or manually.

This paper describes an approach to analyzing, annotating, and evaluating a corpus of scientific articles according to text type categories on a thematic level using XML technology and methods from automatic text categorization.

The text type structure of an article instantiates components, or *topic types* of research papers such as *background*, *researchTopic*, *method*, which are related by a canonical ordering and typical rhetorical relations, all of which constitute characteristic features of the text type, or genre, of scientific articles. Topic types have elsewhere been called *text level categories* (Kando, 1997), or *zones* (Teufel, Carletta, & Moens, 1999).

Building an annotation tool for thematic structure involves automatic classification of segments into topic type categories, thus we additionally provide XML annotations

on other levels of information, namely grammar (syntax and morphology), and logical structure (structural positions according to DocBook markup, cp. Walsh & Muellner, 1999), that can provide features for the classification task. The current aim is thus to examine the correlations between thematic structure and the other levels of analysis to identify linguistic and structural features that constitute topic types. The overall goal of the project SemDoc is to design an empirically based thematic text type ontology that can be used for improved information retrieval/extraction, automatic text summarization and for making scientific articles available to the Semantic Web by automated annotation.¹

In the remainder of this paper, a characterization of our corpus and the methods of analysis and feature extraction using XML annotations on multiple layers as well as automatic text segment classification experiments, will be presented.

2. TEXT TYPE STRUCTURE

Text type schemas representing text-level structure of scientific articles have been devised previously, for instance in the context of automatic text summarization. In Teufel (1999) (see also Teufel et al., 1999), a schema of the seven "argumentative zones" *BACKGROUND*, *OTHER*, *OWN*, *AIM*, *TEXTUAL*, *CONTRAST*, *BASIS* is employed for classifying the sentences of a scientific article and choosing the most suitable sentences for a summary of the article. In Kando (1997), a hierarchical schema with 51 bottom-level text constituent categories is presented that are similar to our topic types discussed below and were used for manually annotating sentences in Japanese research papers. In

¹<http://www.uni-giessen.de/germanistik/ascl/dfg-projekt/>

two experiments, the usefulness of such an annotation in full-length text searching and passage extraction is explored and it is found that it could improve the results. It is also reported that several studies "indicated the feasibility of automatic assignment of categories using surface level natural language processing" (Kando, 1997, p.4). Our text type schema is based on these two approaches but occupies a middle ground between their sizes by including 21 bottom-level topic types, which are supposed to represent the typical structure of texts in the text type of scientific articles. Our aim was to develop an informative schema while sorting out categories we considered primarily functional (like 'Reason for...') and including only purely thematic categories. Moreover, we hypothesized that these 21 topic types could be well distinguished by structural and surface linguistic criteria. The schema is depicted in Figure 1. The edges can be interpreted to represent the *part-of* relation such that a type lower in the hierarchy is a part of the immediately dominating, more global type in terms of text type structure. The order of the categories represents a canonical, expected order of topic types in a scientific article. The text type schema was initially encoded as an XML Schema grammar where topic types are represented by elements that are nested such that the XML structure reflects the structure of the text type structure tree (Figure 2).

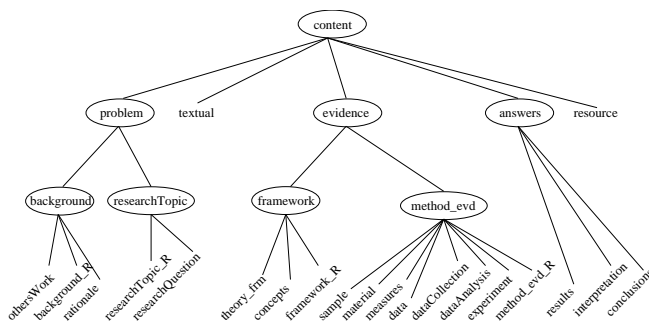


Figure 1: Text type schema

```
<xs:element name="problem">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="background" minOccurs="0">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="othersWork"
              type="xs:string"
              minOccurs="0"/>
            <xs:element name="background_R"
              type="xs:string"
              minOccurs="0"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
  </xs:complexType>
</xs:element>
...
```

Figure 2: XML Schema grammar (extract) for the text type schema

3. DATA COLLECTION

For the analyses and experiments described in this paper, a corpus of 47 research articles from the discipline of linguistics was collected. The articles were taken from the

German online journal 'Linguistik Online'², from the volumes 2000-2003. The articles have an average length of 8639 word forms and deal with subjects as diverse as syntax and morphology, chat analysis, and language learning. To verify the validity of our approach for more than one discipline, we have also collected and analyzed 60 scientific articles from the field of psychology, however, these will not be considered in this report.

3.1. Annotation levels

Following Bayerl, Lungen, Goecke, Witt, and Naber (2003), we distinguish between annotation *levels* and annotation *layers*. An annotation level is a chosen level of information that is initially independent of an annotation design, such as the morphology and syntax levels in linguistics. Annotation layer, in contrast, refers to the realization of an annotation as, for instance, XML markup. There need not be a 1:1-correspondence between annotation levels and layers. We would argue that, for example, in the layer defined by the XHTML DTD, at least one logical level and one layout level are integrated. Conversely, one annotation level may be distributed across several layers. As for the three annotation levels in our setting, one (the structural level) was realized as an independent layer, and two (thematic and grammatical level) were realized in one single annotation layer.

3.1.1. Thematic level

As sketched in section 2., the thematic text type schema represents, amongst other things, an expected canonical order of topic types in a scientific article. Yet, the order of topics in a specific article instance may deviate from it and render an XML instance annotated accordingly invalid. Thus we derive a flat version of the hierarchical XML schema by means of an XSLT style sheet. In the flat XML schema for the thematic annotation layer (called THM), topic types are represented as attribute values of elements called `<group>` and `<segment>`, instead of names of nested elements. The empty `<group>` elements represent topic types that corresponded to the nodes (as opposed to leaves or terminal categories) in the original tree of topic types. The original hierarchical structure is represented via the ID/IDREF attributes `id` and `parent`, similar to O'Donnell's XML representation of rhetorical structure trees (O'Donnell, 2000). Each text segment (a thematic unit, often but not always corresponding to a sentence) is annotated with one terminal topic type, including segments from abstracts, footnotes, or captions. An extract from a THM annotation can be seen in Figure 3.³

The HTML files containing the articles were automatically stripped off their markup, segmented and provided with skeletal markup according to the flat THM schema. Two annotators then had to fill in the attribute values for the topic types of segments using the XML spy editor.

²<http://www.linguistik-online.de/>

³The extract, which is also shown in Figure 4, is taken from Buhlmann (2002)

```

<segment id="s196" parent="g4" topic="results">In den
  Texten ist sehr oft nicht klar, ob ein Maskulinum nur
  auf Männer oder auch auf Frauen referiert.
</segment>
<segment id="s197" parent="g4" topic="interpretation">
  Wichtige Fragen, die die LeserInnen an den Text haben,
  bleiben somit unbeantwortet. Die Politik wird durch den
  fast durchgehenden Gebrauch des generischen Maskulinums
  als "Männersache" dargestellt, Frauen werden, auch wenn
  sie vorhanden sind, selten sichtbar gemacht. Zudem wird
  auch mit geschlechtsspezifisch männlichen Wörtern wie
  Gründerväter der Gedanke an Männer evoziert.
</segment>

```

Figure 3: THM annotation (extract)

```

<sect2>
...
<para POSINFO1="/article[1]/sect1[4]/sect2[4]/para[3]">
  In den Texten ist sehr oft nicht klar, ob ein Maskulinum
  nur auf Männer oder auch auf Frauen referiert. Wichtige
  Fragen, die die LeserInnen an den Text haben, bleiben
  somit unbeantwortet. Die Politik wird durch den fast
  durchgehenden Gebrauch des generischen Maskulinums als
  "Männersache" dargestellt, Frauen werden, auch wenn sie
  vorhanden sind, selten sichtbar gemacht.
</para>
<para POSINFO1="/article[1]/sect1[4]/sect2[4]/para[4]">
  Zudem wird auch mit geschlechtsspezifisch männlichen
  Wörtern wie Gründerväter der Gedanke an Männer evoziert.
</para>
...
</sect2>

```

Figure 4: Annotation according to DocBook (extract)

3.1.2. Structural level

Although the linguistic articles in our corpus are originally provided with HTML markup, HTML itself was not considered as an annotation layer in our scheme, as HTML is known to be a hybrid markup language including a mixture of logical, functional and layout categories. For representing a purer logical structure, the HTML annotations were converted to DocBook markup (Walsh & Muellner, 1999). The DocBook standard was originally designed for technical documentation, but has recently been applied to other scientific writing and is devoid of layout elements such as `font` or `br` in HTML.

We did not employ the whole, very large official DocBook DTD, but designed a new XML schema with a subset of 45 original DocBook elements plus 13 new logical elements not conforming to the DocBook standard which were nevertheless needed for our purposes (for example `tablefootnote`, `toc`, and `numexample`). This reduced XML schema for DocBook was developed in collaboration with HyTex project at the University of Dortmund⁴, to keep the number of admissible elements manageable.

Since the segmentation on the THM layer is into thematic units, whereas the DocBook elements pertain to logical structure units, we did not want to constrain the THM annotation layer to be fully compatible with the DocBook level, requiring that a possible integrated THM-DocBook annotation layer would always yield well-formed XML. Thus the DocBook annotation was realized as a separate XML layer (called DOC).

The annotations were obtained using a perl script that provided the raw DocBook annotations from the HTML

marked up texts, and the XML spy editor for validation and manually filling in DocBook-elements that have no correspondences in HTML. In addition, structural position attributes were added by means of an XSLT stylesheet. These 'POSINFO' attributes make explicit the position of an element in the XML DOM tree of the document instance. The aim is to exploit the position information in the automatic classification of thematic segments in the future. The DocBook annotation of the extract shown in Figure 3 can be seen in Figure 4.

3.1.3. Grammatical level

For an annotation of morphological and syntactic categories to word form tokens in our corpus, the commercial tagger *Machinese Syntax* by Connexor Oy was employed. This tagger is a rule-based, robust syntactic parser available for several languages and based on Constraint Grammar (Karlsson, Voutilainen, & Heikkilä, 1995) and Functional Dependency Grammar (Tapanainen & Järvinen, 1997). It provides morphological, surface syntactic, and functional tags for each word form and a dependency structure for sentences, and besides is able to process and output "simple XML". DTDs for the tag set and for the XML output format are supplied with the software.

Since all annotations provided by *Machinese Syntax* pertain to word forms (dependency structure is realized through ID/IDREF-attributes on word form tags), no conflicts in terms of element overlaps may arise between our THM annotation layer and a potential CNX annotation layer. Speaking in terms of the XML-based multiple layer annotation paradigm (Goecke, Naber, & Witt, 2003), the only meta-relation besides independence that may hold between THM-`<segment>` elements and CNX-tagging elements is *inclusion*. Therefore, the THM and CNX annotations could be integrated into one single annotation layer. This way, not only special tools for the analysis of multiple-layer annotations (Goecke et al., 2003; Bayerl, Lungen, Gut, & Paul, 2003), but also the available query languages for querying information contained in single annotation layers, like XQuery⁵, can be adopted for inferring correlations between topic types and grammatical features. Since the current version of *Machinese Syntax* is able to process only "simple XML", that is, XML without attributes, we implemented two XSLT style sheets, one of which converts our THM-annotations into attribute-free XML by integrating all attribute-value specifications into the names of their respective elements, and another one which reconverts the attribute-free annotations enriched by the CNX-tagging into complex XML.

Out of the large CNX-tag set documented in an extensive manual, we have selected a set of 15 tags (henceforth called CNX-15) that were judged to be valuable for automatic assignment of topic types to text segments of scientific articles. CNX-15 also includes simplified tag specifications that came as a bundle of tags in the original CNX output, cf. the following listing of the CNX-15 tags and their range of values in Table 1.

A third XSLT stylesheet acts as a filter and converter on the integrated THM-CNX annotations to output the THM

⁴<http://www.hytext.info>

⁵<http://www.w3.org/TR/xquery/>

segments plus their CNX-15 tagging in a THM-CNX target format designed for extracting statistics and feature vectors for the automatic classifier.

#	CNX-15 Tag	range of values
1	text	(string)
2	lemma	(lower case string)
3	cmp-head	(lemma of head constituent; lower case string)
4	depend	(dependency category, e.g. loc, dur, frq, i.e. adverbial of location, duration, frequency)
5	pos	N, V, A, ...
6	comparison	POS, SUP
7	nnum	SG, PL (singular or plural of nominal categories)
8	numeral	CARD, ORD
9	pers	SG1, SG2, SG3, PL1, PL2, PL3
10	modal	MODAL (modal auxiliary)
11	fin	INF, IMP, SUBJUNCTIVE, PRES, PAS
12	ncomb	N+
13	unknown	<?>
14	aux	AUX
15	passive	PASS

Table 1: CNX-15 tags derived from the Machine Syntax tag set

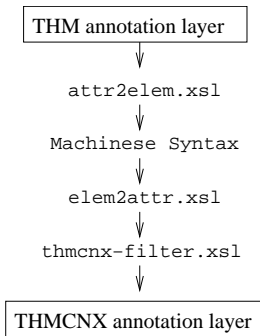


Figure 5: Augmenting THM annotations with grammatical tags

4. AUTOMATIC TEXT SEGMENT CLASSIFICATION EXPERIMENTS

In order to evaluate the feasibility of an automatic annotation of scientific articles according to our THM annotation level introduced in section 3.1.1., we applied different automatic text classification methods to text segments as shown in Figure 3.

The different configurations in the experiments were motivated by the following questions:

- Are thematic structures (of scientific articles) learnable using only general, domain-independent methods?
- Which kind of information (plain text, POS-tag information, morphological analysis, knowledge about the context) has impact on the classification accuracy?
- Which kind of classifier (e.g. KNN, Rocchio) performs best on this task?
- Are there particular topic types which are significantly easier to detect than others?

Our focus has been on the more general question, to which degree thematic structures are learnable, rather than the question how to develop a specialized classifier for the particular task of categorizing text segments according to our text type schema. Thus, in contrast to related work (Teufel, 1999), we restricted ourselves to general and domain-independent classification and pre-processing methods, which are, in principle, also applicable to any other kind of text.

In contrast to standard text classification tasks (Sebastiani, 2002, cf.), the pieces of text in our experiments are much smaller than ordinary documents, and sometimes consist only of a single word or phrase as in the case of headlines. On the other hand, the text segments to be classified appear in a context, which is an additional source of information, not available in case of the standard problem of document categorization.

4.1. Vector Representation

The classification experiments have been carried out at different levels of linguistic analysis:

- inflected word forms (from the raw text)
- stems (from the lemma annotation as shown in Table 1)
- part-of-speech patterns (from the pos annotation as shown in Table 1)
- head-lemma (the cmp-head annotation as shown in Table 1)

Some of these levels of description have also been used in combination (e.g. part-of-speech patterns combined with head lemmata).

For the purpose of our classification experiments each text segment has been represented as a (sparse) probability distribution vector of its units (*units* are e.g. inflected words, POS tags). Feature vectors have been generated directly from the THM-CNX annotation layer introduced above. We did not use TF*IDF or other feature weighting methods, but the bare probability of a term, given its segment⁶. Neither did we employ stop word lists or frequency-based filtering in order to reduce the feature space.

⁶TF*IDF weighting had been used in preliminary experiments, but did not improve the results.

4.2. KNN Classification

The basic idea of the K nearest neighbor (KNN) classification algorithm is to use already categorized examples from a training set in order to assign a category to a new object. The first step is to choose the K nearest neighbors (i.e. the K most similar objects according to some similarity metric, such as cosine) from the trainings set. In a second step the categorial information of the nearest neighbors is combined, in the simplest case, by determining the majority class.

The version of KNN classification, adopted here, uses the *Jensen-Shannon divergence* (also known as *information radius* or *iRad*) as a (dis-)similarity metric:

$$\text{iRad}(q, r) = \frac{1}{2}[D(q \parallel \frac{q+r}{2}) + D(r \parallel \frac{q+r}{2})]$$

where $D(x \parallel y)$ is the Kullback-Leibler divergence (KL divergence) of probability distributions x and y :

$$D(x \parallel y) = \sum_{i=1}^n x(i)(\log(x(i)) - \log(y(i)))$$

iRad ranges from 0 (identity) to $2\log 2$ (no similarity) and requires that the compared objects are probability distributions.

Let be $N_{O,C} = \{n_1, \dots, n_m\}$ ($0 \leq m \leq K$) the set of those objects among the K nearest neighbors of some new object O that belong to a particular category C . Then the score assigned to the classification $O \in C$ is

$$\text{score}(O, C) = \sum_{j=1}^m \text{iRad}(O, n_j)^E.$$

Depending on the choice of E , one yields either a simple majority decision (if $E = 0$), a linear weighting of the iRad similarity (if $E = 1$), or a stronger emphasis on closer training examples (if $E > 1$). Actually, it turned out that very high values of E improved the classification accuracy. Finally, the KNN scores for each segment were normalized to probability distributions, in order to get comparable results for different K and E , when the KNN classifications get combined with the bigram model (see section 4.3., below).

4.3. Bigram Model

The bigram model gives the conditional probability of a topic type T_{n+1} , given its predecessor T_n .

For a sequence of segments $s_1 \dots s_m$ the total score $\tau(T, s_i)$ for the assignment of a topic type T to s_i is the product of the bigram probability, given the putative predecessor topic type (i.e. the topic type T' with the highest $\tau(T', s_{i-1})$ computed in the previous step), and the normalized score of the KNN classifier. The total score of the topic type sequence is the product of its τ scores.

4.4. Training and Evaluation

For the classification experiments we used the data collection described in section 3.. For each test document the bigram model and the classifier were trained with all other documents. The overall size of the data collection

was 47 documents. Thus, each classifier and each bigram model has been trained on the basis of 46 documents, respectively. The total number of segments was 7330. 23 different classes⁷ have been manually assigned to the segments of the sample⁸. The number of features varied by the respective choice of data representation: The total number of stems was 33,000 (about 400,000 tokens), the total number of POS tag types was 14.

Additional experiments have been carried out using a simplified Rocchio classifier. This classifier computes the centroid vector for each class and assigns the category of the centroid vector that has the least iRad distance relative to the segment in question.

4.5. Results

We performed several hundred classification tests with different combinations of data representation, classification algorithm, and classifier parameter setting. Table 2 summarizes some results of these experiments, table 3 shows the precision and recall values of the K-13-E-40 classifier with bigram model (last line in table 2) for each topic type. For illustrative purpose, we also included a configuration, where all other segments (i.e. including those from the same document) were available as training segments ('KNN*' in the butlast line of table 2).

classifier	data	K	E	accuracy classifier	accuracy classifier + bigram
KNN	head	13	45	39.433	42.050
KNN	POS	20	40	40.328	41.751
KNN	stem	17	45	38.959	41.196
Rocchio	POS+head	-	1	36.099	20.876
KNN*	POS+head	13	40	54.416	-
KNN	POS+head	13	40	43.812	45.872

Table 2: Results

The standard deviation across topic types of about 24 (both for recall and precision) indicates that the "learnability" of topic types differs enormously. The topic type `resource` has been learned almost perfectly, while other topic types (e.g. `material`) have no recall, at all.

4.6. Discussion

The data collection used for the classification experiments is restricted in many respects: one language (German), one type of document (scientific article), one thematic domain (linguistics), one thematic ontology, and only 46 training documents. Thus, the results of our experiments can only give a rough idea of the lower bound of the accuracy that can be achieved by the application of general,

⁷The classes `void_C` and `void_meta` in Table 2 were non-thematic labels assigned to incomplete segments and metadata in the corpus (such as author, affiliation, and acknowledgements), respectively. Thus, they are not part of the abstract thematic schema depicted in Figure 1.

⁸The number of training examples per class ranges from 5 (`experiment`) to 1643 (`resource`). 3 classes have less than 10 training examples.

class	recall	precision
background_R	16.346	23.944
concepts	1.639	5.770
conclusions	29.602	25.813
data	6.195	25.000
dataAnalysis	0.442	3.846
dataCollection	0.000	0.000
experiment	0.000	0.000
framework_R	30.914	23.842
interpretation	15.209	21.277
material	0.000	0.000
measures	0.000	0.000
method_evd_R	5.556	40.000
othersWork	72.673	31.311
rationale	0.000	0.000
researchQuestion	23.296	75.926
researchTopic_R	34.163	45.619
resource	97.018	93.490
results	27.343	24.895
sample	0.000	0.000
textual	29.750	40.067
theory_frm	0.000	0.000
void_C	0.000	0.000
void_meta	67.083	83.420

Table 3: Recall and precision

domain-independent classification methods to this particular kind of document. The upper bound (e.g. if larger training sets are available) still remains unclear. Additionally, the classification experiments reported in this paper are, to our knowledge, the first attempt to apply domain-independent machine learning methods to the problem of identifying the topic types of text segments. Because of the novelty of the approach, there are no "baseline" results that can serve as a standard.

Besides the limitations, stated above, there are some interesting results:

- The accuracy of the best configuration is close to 50%
- The choice of the classification algorithm seems to play an important role (Rocchio vs. KNN).
- The POS-tag distribution of text segments turned out to be nearly as informative as the "bag-of-words" representation.
- The usage of a bigram model improved the accuracy results in almost all configurations.
- The variance of classification accuracy across topic types is extremely high.

5. CONCLUSION AND PROSPECTS

Many applications, e.g. in the context of the Semantic Web, require rich and fine-grained annotations on linguistic levels. In this paper we presented a multiple layer approach to the semantic, grammatical and structural annotation of scientific articles. We carried out experiments on automated annotation of text segments with topic types, using

general and domain-independent machine learning methods. We achieved an average accuracy of almost 50%. Although the results probably suffer from limitations of our data collection (small sample size, restricted thematic domain), our main conclusion is that at least some of the topic types of our hierarchy are successfully learnable. Other classification algorithms (e.g. support vector machines), feature selections methods, and/or larger training sets may yield further improvements. Our future work will focus on the integration of structural position information from the DOC annotation layer, usage of additional information from deep syntactic analyses, and the question to which degree our results are generalizable to other thematic domains and languages.

References

- Bayerl, P. S., Längen, H., Goecke, D., Witt, A., & Naber, D. (2003). Methods for the semantic analysis of document markup. In *Proceedings of the ACM symposium on document engineering (DocEng 2003)*. Grenoble.
- Bayerl, P. S., Längen, H., Gut, U., & Paul, K. (2003). Methodology for reliable schema development and evaluation of manual annotations. In *Workshop notes for the workshop on knowledge markup and semantic annotation, second international conference on knowledge capture (K-CAP 2003)* (p. 17-23). Sanibel, Florida.
- Bühlmann, R. (2002). Ehefrau Vreni haucht ihm ins Ohr... Untersuchungen zur geschlechtergerechten Sprache und zur Darstellung von Frauen in Deutschschweizer Tageszeitungen. *Linguistik Online, 11*. (<http://www.linguistik-online.de>)
- Giles, C. L., Bollacker, K., & Lawrence, S. (1998). Cite-Seer: An automatic citation indexing system. In I. Witten, R. Akscyn, & F. M. Shipman III (Eds.), *Digital libraries 98 - the third ACM conference on digital libraries* (pp. 89-98). Pittsburgh, PA: ACM Press.
- Goecke, D., Naber, D., & Witt, A. (2003). Query von Multiebenen-annotierten XML-Dokumenten mit Prolog. In U. Seewald-Heeg (Ed.), *Sprachtechnologie für die multilinguale Kommunikation. Beiträge der GLDV-Frühjahrstagung, Köthen 2003* (Vol. 5, p. 391-405). Sankt Augustin: gardez!-Verlag.
- Kando, N. (1997). Text-level structure of research papers: Implications for text-based information processing systems. In *Proceedings of the British computer society annual colloquium of information retrieval research* (p. 68-81).
- Karlsson, F., Voutilainen, A., & Heikkilä, J. (Eds.). (1995). *Constraint grammar: a language-independent system for parsing unrestricted text* (Vol. 4). Berlin and N.Y.: Mouton de Gruyter.
- O'Donnell, M. (2000). RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proceedings of*

the international natural language generation conference (INLG'2000) (pp. 253 – 256). Mitzpe Ramon, Israel.

- Sebastiani, F. (2001). Organizing and using digital libraries by automated text categorization. In L. Bordonì & G. Semeraro (Eds.), *Proceedings of the AI*IA workshop on artificial intelligence for cultural heritage and digital libraries* (p. 93-94). Bari, Italy.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47.
- Tapanainen, P., & Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the 5th conference on applied natural language processing* (p. 64-71). Washington D.C.
- Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text*. Unpublished doctoral dissertation, University of Edinburgh.
- Teufel, S., Carletta, J., & Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*.
- Walsh, N., & Muellner, L. (1999). *DocBook: The definitive guide*. O'Reilly.

ACKNOWLEDGEMENT

This work was supported by the German Research Foundation (DFG) in the context of research group nr. 437 *Texttechnologische Informationsmodellierung*.