

# **Towards Automatic Annotation of Text Type Structure**

## **Experiments Using an XML-Annotated Corpus and Automatic Text Classification Methods**

**Hagen Langer**

Justus-Liebig-University, Giessen  
University of Osnabrück

**Harald Lungen**

Justus-Liebig-University,  
Giessen

**Petra Saskia Bayerl**

Justus-Liebig-University,  
Giessen

# Outline

- Project description and purposes
  - Corpus Data
  - Annotation levels
- Experiment: Automatic Text Segment Classification
- Conclusions and Prospects

# Project Description: SemDoc

- *SemDoc* wants to describe the semantic structure of text types
- Example domain is `scientific article` with a restriction to linguistics and psychology
- Overall goal is the design of an empirically based text type ontology
  - information retrieval and extraction
  - automatic text summarization
  - automatic annotation for the semantic web

# What did Author X hypothesize?

upload files to the bulletin board before they were able to download. In response to this requirement users would upload their own text files and artwork or randomly generated text files and would be able to download high quality content generated by others. In the experiments described below we address both kinds of free riding.

## Experiments

In the following section we describe the experiments used to test the following three hypotheses:

- Hypothesis 1: A significant portion of Gnutella peers are free riders.
- Hypothesis 2: Free riders are distributed evenly across different domains (and by speed of their network connections).
- Hypothesis 3: Peers that provide files for download are not necessarily those from which files are downloaded.

## Measuring downloads

One of the features that attract users to Gnutella is the difficulty in associating queries to any particular peer/user. Given a query message it is virtually impossible (unless some large percentage of peers collude) to find the peer that originated the query. The unfortunate side effect of this property is to make it impossible to experimentally measure the number of queries and files downloaded by each client. This forces us to make assumptions about downloads in order to measure them.

One possible assumption is that users that share a high number of files had to have downloaded them, so those that share more also download more. In this case, there is no free riding. The other possible assumption is that users who have no files are those that will try to access them. Therefore the fewer files a user has the more likely he is to download them, resulting in rampant free riding.

Since we unfortunately have no way of knowing which of these two extremes is closest to reality, we assume that the truth is somewhere in between.

## Experimental Setup

In order to perform monitoring experiments on the Gnutella network it was necessary to modify a Gnutella client to log messages flowing through the system. We elected to use the Java based Furi client [Fu00] which was a full featured implementation, with numerous hooks for logging.

The Furi client was then executed for a 24-hour period over a weekend in August of 2000 (Saturday 1pm to Sunday 1pm) [1]. During this time period we collected both pong and query response messages from normal Gnutella users. A shorter trace during a weekday shows results consistent with the weekend findings. In the 24-hour period we observed 35,352 hosts issuing ping messages, which shared a total of 3,304,046 files.

# How high are response rates in student questionnaire studies?

**AGR BUSINESS MANAGEMENT**

**THE MANAGER UNDER STRESS**

A few nights ago I was watching the evening news on TV. This particular network had pulled together a brief news report on the impact that rising unemployment and recession were having on the "average American's paycheck." This network chose the hypothesis that economic instability and recession were mirrored in the minds of men via emotional instability and mental depression. To test this hypothesis, network reporters conducted a series of random interviews with corporate executives, industrial psychologists, and some of the more recent additions to the ranks of the unemployed. While the interview process and the results obtained could hardly be categorized as "scientific," there was little doubt in the mind of this TV observer that stress in the business community was positively correlated with the Dow-Jones Index.

Stress, and the emotional havoc it can cause, has always been as much a part of the modern agribusniss industry as the annual balance sheet and profit and loss statement. It affects the younger person who is striving to advance up the management ladder. It affects the middle-aged executive who confronts, for the first time, the possibility that his professional career may have reached its peak. And, of course, it does not leave unaffected the person who has already made it to the top, but who must now learn to survive in an economy which no longer abides by old rules and practices, and which seems to be running at a pace almost out of control. When considering the usual social and family affairs plus the uncertainties and pressures of his business, stress is virtually unavoidable for the agribusniss manager.

Regarding the latter item, 1975 was extremely difficult for the average agribusniss manager. Supplies were often difficult to obtain. If they could be obtained, each shipment was accompanied by a revised price list. Processors were forced to pay record high prices for many row crops and most feed grains. Ever-rising fuel prices had a particularly devastating effect on the transportation sector. And finally, the livestock industry was on the skids due to the rapid rise in feed costs, accompanied with beef prices at their farm-gate snapping to one-half that of just a year earlier. No sector of the industry was left unaffected, and 1974 was truly a year of high stress, i.e., a real ulcer creator.

**How Do You Identify Stress?**

The above suggests that stress is an unfortunate, but unavoidable, part of a manager's life. And, it has been suggested that the year 1974 was generally conducive to more stress than usual. In view of the above, there is little the modern agribusniss manager can do to escape from stress. What is important, then, is that: 1) he is able to identify stress when he confronts it, and 2) he is or is willing to develop the ability to cope with it. These are important because the very personality traits that are indicative of a successful manager -- his drive, competitiveness, and an emphasis on individual strength and superior abilities -- make it extremely difficult for him to recognize stress before it gets too great and even harder for him to seek assistance when it is required.

1

WASHINGTON STATE UNIVERSITY & U.S. DEPARTMENT OF AGRICULTURE COOPERATING

College Opinion Survey sample group size consisted of 3000 and the Gas Company Opinion Survey was mailed to 1000 customers.

The chi-square goodness-of-fit test was used to determine if there was a relationship between response rates and sponsorship. The hypothesis under test was that the response rates for each cover letter treatment group did not differ. The chi-square test of independence was employed to determine if response rates by sponsorship and gender were independent.

**III. RESULTS AND ANALYSIS**

The overall customer response rate to the surveys was 41.6 percent based on 1664 responses with no follow-up appeals. The response rate to the college student surveys sent to 3000 of the customers was 39.3 percent (1178). The response rate to the Gas Company survey sent to 1000 customers was 48.6 percent (486).

The expected proportion of respondents to the Student Survey was 75 percent (1248), 71 percent (1178) actually responded. The expected proportion of respondents to the Gas Company Survey was 25 percent (416), 19 percent (486) actually responded. The difference in response rates between cover letter treatments is statistically significant at  $\alpha = .001$  as Table 1 illustrates.

**Table 1. Cover Letter Responses**

Cover Letter	Questionnaires	Total Sample
College		3,000 [75%]
Received	1178 [70.8%]	
Expected	1248 [75%]	
	$\chi^2 = 3.926$	
Gas Company		1000 [25%]
Received	486 [29.2%]	
Expected	416 [25%]	
	$\chi^2 = 11.779$	
Total	1664 [41.6%]	4000
	$DF = \chi^2_c = 15.7$	$\alpha = .001$
	$\chi^2_c = 10.8$	

The response rates by cover letter treatment between the genders were also examined. Fifty-six percent of the responses were male and 44 percent were female. The expected percentage of male and female responses to the College cover letter was 71.9 percent. Males seemed to prefer the student involvement with 74 percent of the total male returns (796) responding to this appeal while females were

# What are research questions concerning stress in managers?

# Methods

- Corpus compilation of scientific articles
- Annotation of text type specific knowledge on different levels
- Experiment on automatic annotation of text type structure information

# Corpus Compilation

- Compilation of a corpus of scientific articles
  - Linguistics: 47 German articles
  - Psychology: 60 English articles
- Sources:
  - Linguistic articles: „Linguistik Online“
  - Psychological articles: 3 articles per journal (ISI-Ranking)

# Corpus Annotation

- XML Annotation of text type specific information on different levels
  1. Thematic information
  2. Structural information
  3. Morphological and syntactic information
  4. [Rhetorical information]
- Annotation format is XML
- Annotations are done
  - Manually: thematic and structural level
  - Automatically: syntactic level

# 1. Thematic Annotation Level

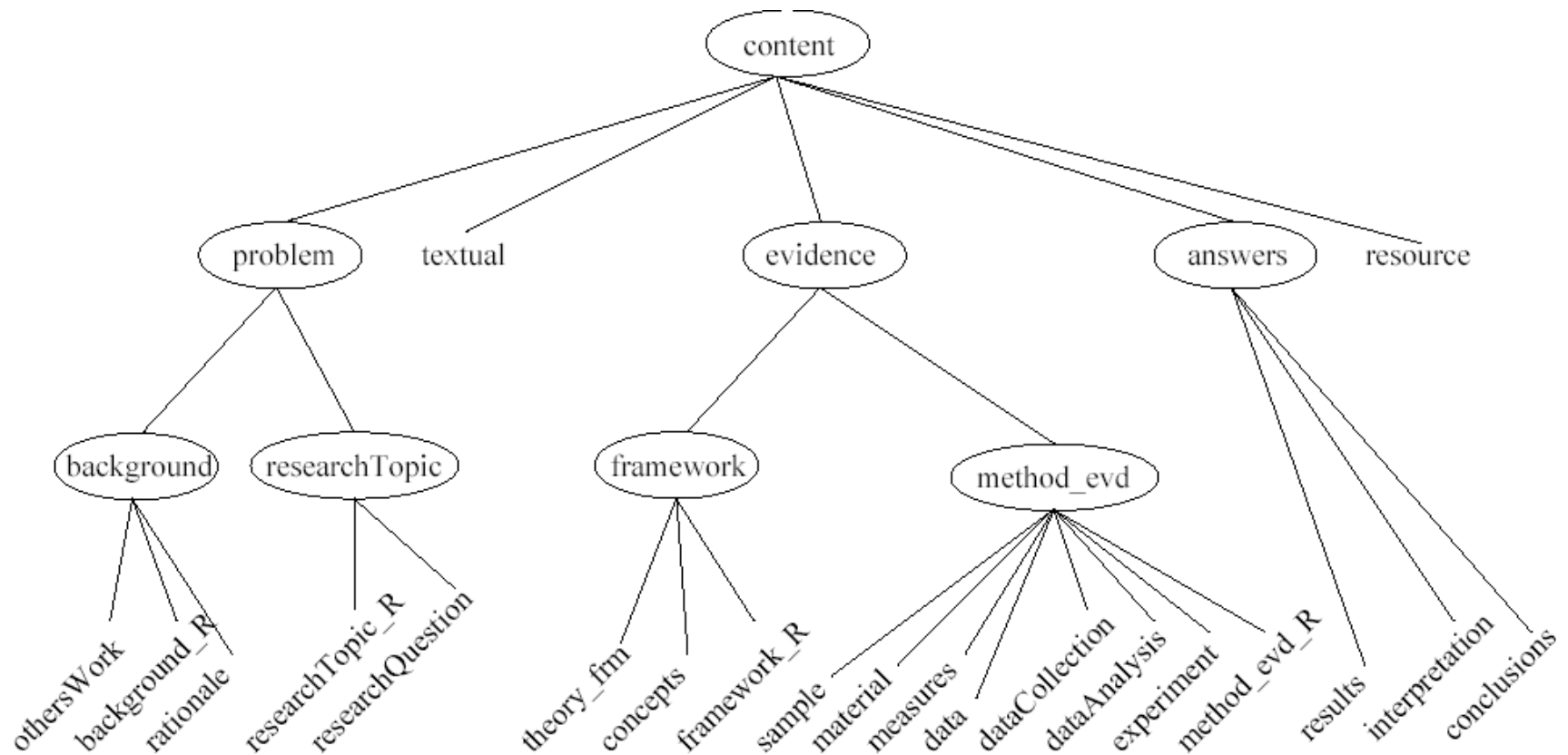
- Thematic annotations describe what a text segment „*is about*“

```
topic = "hypothesis"
```

```
topic = "results"
```

- 21 such topics are provided in a hierarchically ordered schema

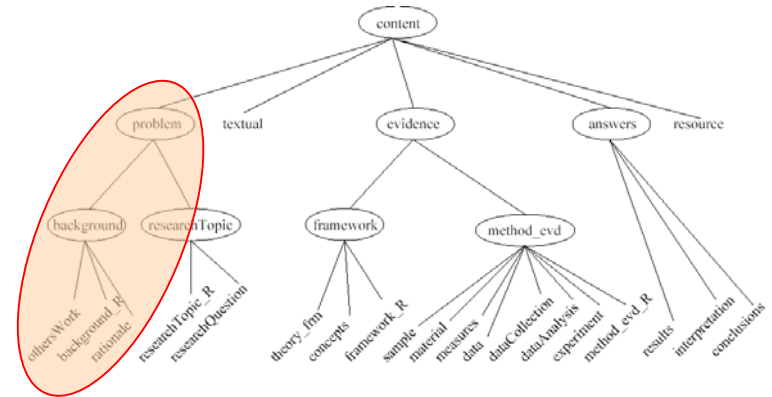
# Thematical Text Type Schema



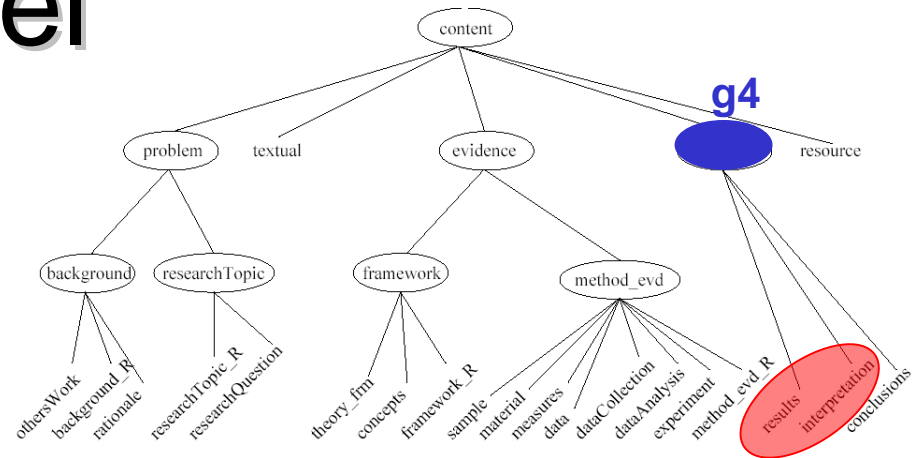
# Thematic Representation

```
<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>
```

...



# Thematic Level



<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. [...]

</segment>

## 2. Structural Annotation Level

- ‚Light‘-version of DocBook with
  - 45 DocBook-elements
  - 13 new logical elements, e.g. `tablefootnote`, `toc`, `numexample`
- Additional POSINFO attributes denote the position of an element

```
POSINFO1="/article[1]/sect1[4]/sect2[4]/para[3]
```

# Structural Annotation

```
<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>
```

# 3. Syntactic Annotation

- Automatic annotation using *Machinese Syntax* (Connexor Oy)
- Morphological, surface syntactic, and functional tags for word forms

# Syntactic Annotation Tags

#	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

Selection of 15 tags  
(CNX-15)

# THMCNX-Layer

- Integration of syntactic (CNX) and thematic (THM) levels

```
<segment id="s14" parent="g2" topic="researchTopic">
  <token id="w405">
    <text>Das</text>
    <lemma>das</lemma>
    <depend head="w406">det</depend>
    <tags>
      <syntax>@PREMOD</syntax>
      <morpho>DET Def NEU SG NOM</morpho>
    </tags>
  </token>
  ...
</segment>
```

# Automatic Text Segment Classification Experiment

# Purpose

Evaluation of the feasibility of an automatic annotation with respect to the *thematic level*:

1. Are thematic structures learnable using only general, domain-independent methods?
2. Which kind of information has impact on the classification accuracy?
3. Which kind of classifier performs best on this task?
4. Are there particular topic types which are easier to detect than others?

# Methods

- Corpus: 47 German linguistic articles
- Classification task:
  - Domain-independent classification and preprocessing methods
- Very specific topic that do not concern whole documents, but (small) text passages
  - Units, i.e. segments can contain sentences, or clauses

# Classification Features

- Inflected word forms (from raw text)
- Stems (`lemma` annotation in CNX-layer)
- Part-of-speech patterns (`pos` annotation in CNX-layer)
- Head-lemma (`cmp-head` annotation in CNX-layer)
- Combinations of these features

# Vector Representation

- Text segments are represented as *vectors* containing component such as pos tags, inflected words
- Feature vectors were directly generated from the THMCNX-layer
- Implementation was done
  - Without weighting of features, e.g. TF\*IDF
  - Without stop lists and frequency filtering

# Classification Algorithms

- K-nearest-neighbor (KNN) Classification using Jensen-Shannon divergence (iRad) as similarity metric

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

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

# Bigram Model

- Description of the probability of a topic  $T_{n+1}$  given its predecessor topic  $T_n$
- For a sequence of segments a total score for each topic  $T$  is calculated according to the bigram probability and the KNN classifier score.

# Training and Evaluation

- For each test document the bigram model and the classifier were trained with all other documents, i.e. 46 training texts or 7330 test segments
- A simplified Roccio classifier was used as a standard of comparison

# Training and Evaluation

- Classification tests with different combinations of
  - data representation
  - classification algorithm, and
  - classifier parameter setting
- For each topic type precision and recall were calculated

# Results

- Accuracy of the best configuration is close to 50% (baseline 22%)
- KNN is significantly better than the simplified Roccio algorithm
- The bigram model improves accuracy in most configurations
- Variance of classification accuracy across topics is very high

# Recall and Precision [1]

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



# Recall and Precision [2]

class	recall	precision
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



# Discussion

- Automatic thematic annotation of small text units is possible, but accuracy differs dramatically for different topic types
- Some topics were underrepresented, e.g. only 5 examples for topic `experiment`
- However, results are not well generalizable, since the data basis is restricted according to size, language, document type, and domain

# Conclusions and Prospects

- A multiple layer approach of semantic, grammatical, and structural annotations of scientific articles seems promising
- Further work is needed to improve automatic annotation accuracy:
  - Integration of structural position information
  - Using additional syntactic information
  - Testing other domains and languages

Diagram illustrating the text "SemDoc" with three red boxes highlighting specific characters. The first box highlights the 'S', the second highlights the 'e', and the third highlights the 'c'. Arrows point from above to the 'S' box and from below to the 'c' box. A grey box highlights the 'm' character, with an arrow pointing from the 'e' box to it.