

BA: Structured Extraction of Terms and Conditions from German and English Online Shops

Tobias Schamel, 6th Sep 2021, Final Presentation (Advanced Seminar)

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

1. Motivation

2. Research Question

3. Content Extraction

- Requirements
- Implementation
- Processing Demo

4. Hierarchy Extraction

- Requirements
- Implementation
- Processing Demo

5. Architecture & Additional Technology

6. Evaluation

- Content Extraction
- Hierarchy Extraction

7. Conclusion & Future Work

Whom of you did ever read the Terms and Conditions when buying products online?

- no standard solutions for parts of the NLP pipe (structured extraction from websites)
- structured extraction is needed in legal documents to allow the drawing of conclusions later in the pipe
 - no existing solutions for structured & hierarchical extraction of T&C

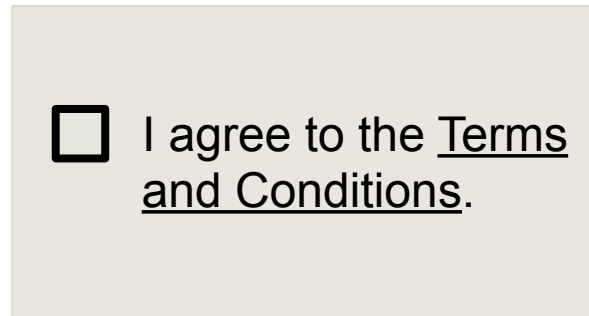


Figure 1.1: Motivation.



0.1% to 0.2% of customers
read T&C in online shops

[Journal of Legal Studies]Bakos, Y., Marotta-Wurgler, F., Trossen, D.: Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts.

1. Motivation

2. Research Question

3. Content Extraction

- Requirements
- Implementation
- Processing Demo

4. Hierarchy Extraction

- Requirements
- Implementation
- Processing Demo

5. Architecture & Additional Technology

6. Evaluation

- Content Extraction
- Hierarchy Extraction

7. Conclusion & Future Work

- RQ1: What are special requirements for the extraction of contracts in comparison with *regular* (i.e. news articles, blogposts) content?
- RQ2: How to extract the relevant parts from contracts in raw HTML?
- RQ3: How to extract the structure and the hierarchy of paragraphs, (sub-)titles, and related clauses?

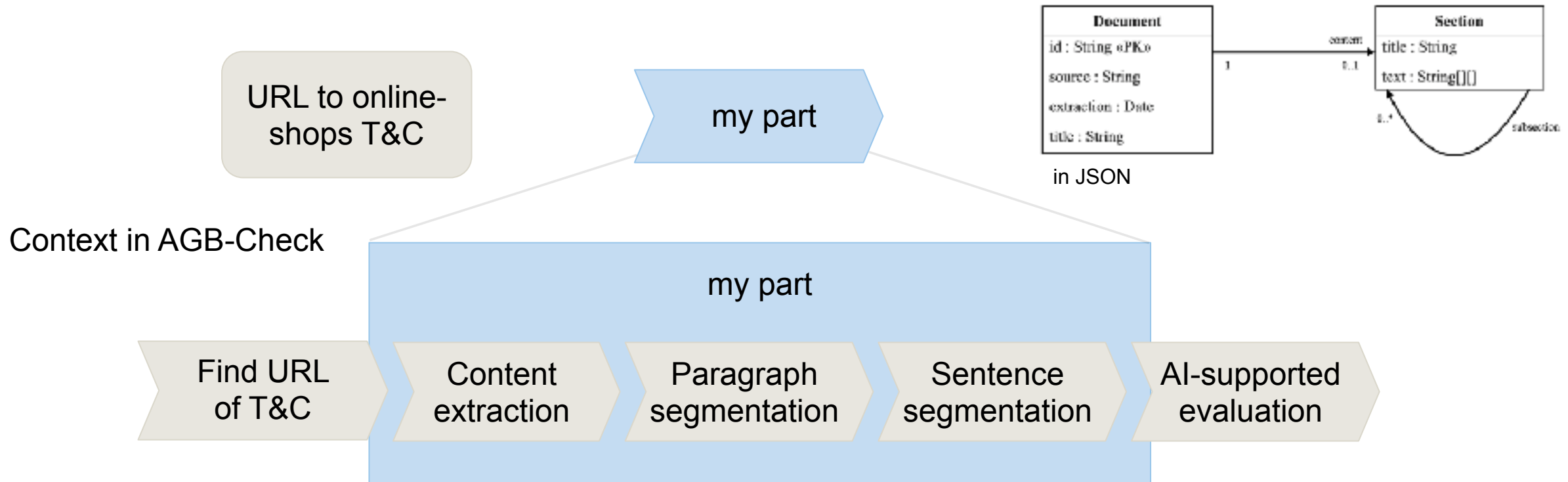


Figure 2.1: Context in AGB-Check.

Example

FAST DELIVERY

BUYER PROTECTION

PRIVACY

Order status

Irrelevant content

Tyres+ Complete Wheels+ Wheels+ Speciality & Truck Tyres+ Motorcycle Tyres+

Home > General Terms and Conditions as per 5th of April 2007

General Terms and Conditions as per 5th of April 2007

§ 1 Conclusion of the contract/order process

Via our Internet portals you may order the listed items using an Internet form or via fax and telephone respectively. By providing the details as required as part of the order process and by sending this order in you submit binding offer; in this connection the order via telephone, fax or Internet have the same legal validity. By ordering you firmly declare to accept these general terms and conditions and that you wish to acquire the ordered goods. We will confirm your incoming order with a formal acknowledgement of receipt; this acknowledgement of receipt is not an acceptance of your order. The contract takes effect from receipt of the goods with your approval, at the latest, however, after the expiry of the return period of 14 days.

§ 2 General conditions of delivery and payments

For any information regarding the costs for packaging and posting as well as the payment conditions, please, refer to the conditions of delivery and payments of the respective Internet portal. For deliveries abroad we reserve the right to possibly limit the payment options.

§ 3 Cancellation Policy

You may revoke your contractual declaration within two weeks without providing reasons in written form (e.g. fax, email). The period commences with the receipt of the goods at the earliest. To ensure the period of revocation the timely sending of the revocation or the object suffices.

The revocation shall be addressed to:
Email: mytyres@delti.com or
Fax: 01212-103310 or
Address: DeltiLog Ltd., 6 Langcote Court, Market Square, Witney, Oxfordshire, OX29 6FG

Figure 2.2: T&Cs of www.tyres-guru.co.uk, 210905.



```
{
  "content": [
    {
      "subsections": [
        {
          "subsections": [],
          "text": [
            [
              "Via",
              "our",
              "Internet",
              "portals",
              "you",
              "may",
              "order",
              "the",
              "listed",
              "items",
              "using",
              "an",
              "Internet",
              "form",
              "or",
              "via",
              "fax",
              "and",
              "telephone",
              "respectively",
              "."
            ],
            "title": "\u00a7 10 Applicable law"
          ],
          "text": [
            [
              "General Terms and Conditions as per 5th of April 2007"
            ],
            "extractionDate": [
              816109,
              39,
              31,
              19,
              5,
              9,
              2021
            ],
            "id": 4803567503487537073,
            "source": "https://www.tyres-guru.co.uk/AGBs.html",
            "title": "General Terms and Conditions as per 5th of April 2007 - Tyre Guru"
          ],
          "by": [
            "By",
            "providing",
            "the",
            "details",
            "as",
            "required",
            "as",
            "part",
            "of",
            "the",
            "order",
            "process",
            "and",
            "by"
          ]
        }
      ]
    }
  ]
}
```

Code 2.1: Excerpt of parsed JSON, start & end.

1. Motivation

2. Research Question

3. Content Extraction

- Requirements
- Implementation
- Processing Demo

4. Hierarchy Extraction

- Requirements
- Implementation
- Processing Demo

5. Architecture & Additional Technology

6. Evaluation

- Content Extraction
- Hierarchy Extraction

7. Conclusion & Future Work

Evaluation of Existing Solutions

boilerpipe (Canola, LargestContent & Article), trafilatura, jusText

Methodology of Existing Solutions:

- Shallow Text Features to classify text as either *Content* or *NotContent*
- identification of largest container
- relative position of DOM-node

Problems:

- only extracts one of two large sections
- addresses, phone numbers, etc.

Manual Review

Potential Problems:

- withdrawal forms (different style)

Properties of the Content:

- main content fills largest part of the page
- same style for the main content
- content is not interrupted (continuous)

Identify Most Common Style
(MCS)

Find *Lowest Common Ancestor*
holding a minimum of 85% of the characters in the MCS

Fallback Solution
find the largest subsequence of direct body-children
holding the MCS

Rationale:

- uninterrupted continuous content
 - DOM-node or series of subsequent DOM-nodes form main content
- main content text usually shares a common visual style (and depth)
 - may also include different styles for headlines or withdrawal forms
- use of CMS (Content Management System)
 - main-content often grouped in a container (i.e. a `<div>`)

Identify Most Common Style
(MCS)

Find *Lowest Common Ancestor*
holding a minimum of 85% of the characters in the MCS

Fallback Solution
find the largest subsequence of direct body-children
holding the MCS

1. count number of characters for each style
 - naive style and short text exclusion
2. determine the style with the highest amount of characters (MCS)
 - style approximations
 - naive style: `<tag>${<attributes>}`
 - rendered style (see Figure 3.1)
 - additional rules
 - short text exclusion: only account for nodes containing more than 3 words

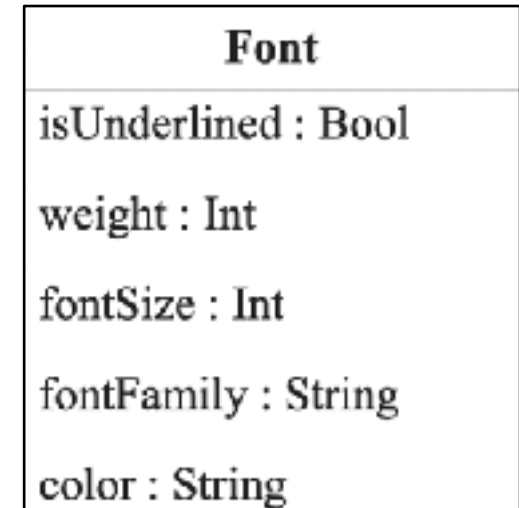


Figure 3.1: UML representation of fonts used to represent rendered styles.

Identify Most Common Style
(MCS)

Find *Lowest Common Ancestor*
holding a minimum of 85% of the characters in the MCS

Fallback Solution
find the largest subsequence of direct body-children
holding the MCS

1. search for the lowest node (highest depth) holding at least 85% of the characters of the MCS
2. if such a node is available, it is the main content node; if not, the fallback solution is used

- threshold 85%:
 - a higher threshold would extract too much content or trigger the fallback in more cases, worsening the overall content extraction performance
 - a lower threshold would extract too little content, i.e. a node holding only a part of the main content in more cases, worsening the overall content extraction performance

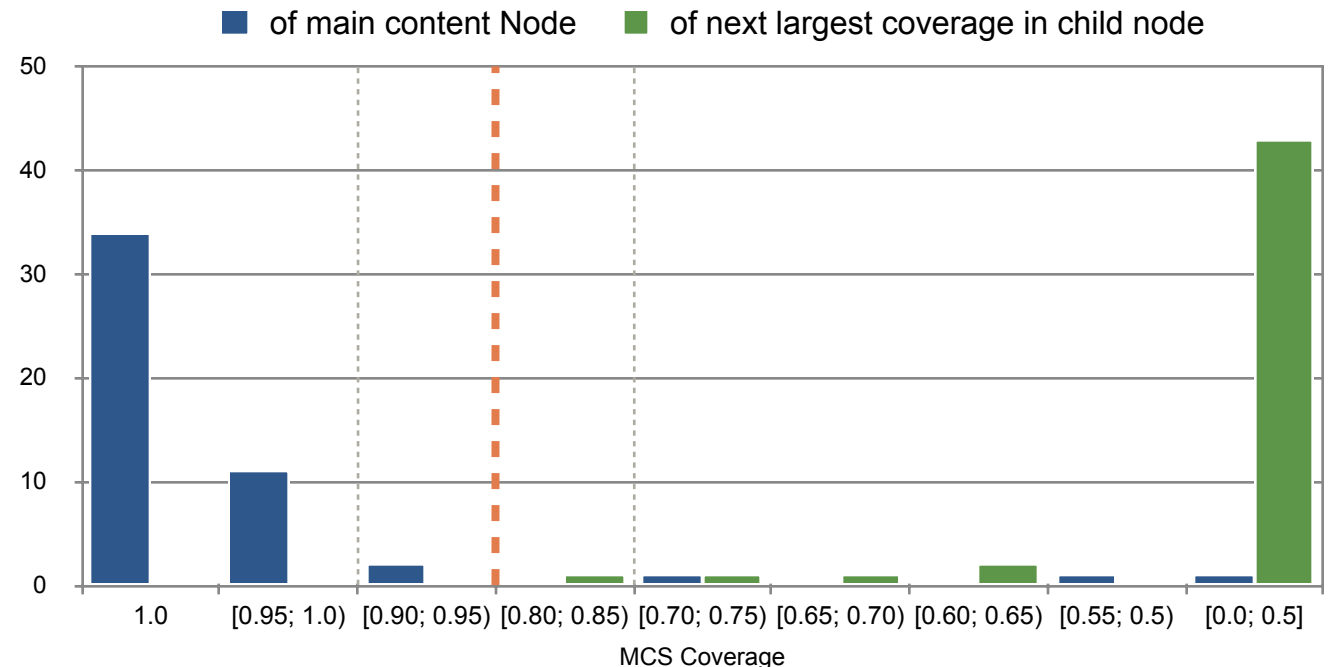


Figure 3.2: Amount of nodes covering a specific amount of the MCS characters*.

* Only existing ranges are included in the diagram.

Identify Most Common Style
(MCS)

Find *Lowest Common Ancestor*
holding a minimum of 85% of the characters in the MCS

Fallback Solution
find the largest subsequence of direct body-children
holding the MCS

1. no node covering at holding at least 85% of MCS characters found by *LowestCommonAncestor Extractor*
 2. Find largest subsequence of direct body-children holding the MCS
- assumption:
 - there is no container holding the main content
 - main content spread over multiple direct body-children
 - risk:
 - there is a container holding the main content but it is made of less than 85% of the MCS characters

Raw HTML

```
<html>
  <head>
    <title>Terms and Conditions of Demo-Shop</title>
  </head>
  <body>
    <div>
      <h1>Welcome to the Demo-Shop</h1>
      <table border="solid">
        <tbody><tr>
          <td>Navigation</td><td>Products</td>
          <td>Sale</td><td>About Us</td>
        </tr></tbody>
      </table>
    </div>
    <div>
      <h2>Terms and Conditions</h2>
      <h3>1. Lorem Ipsum</h3>
      <p>dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.</p>
      <h4>1.1 Donec quam</h4>
      <p>felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, <b>aliquet nec</b>, vulputate eget, arcu.</p>
      <h4>1.2 In enim justo, rhoncus</h4>
      <p>ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.</p>
      <h3>2. Aenean leo</h3>
      <p>ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.</p>
    </div>
    <p>Thanks for visiting Demo-Shop</p>
  </body>
</html>
```

Code 3.1: Raw HTML for the Demo-Shop used in the processing demo.

Welcome to the Demo-Shop

Navigation	Products	Sale	About Us
------------	----------	------	----------

Terms and Conditions

1. Lorem Ipsum

dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

1.1 Donec quam

felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, **aliquet nec**, vulputate eget, arcu.

1.2 In enim justo, rhoncus

ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.

2. Aenean leo

ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.

Thanks for visiting Demo-Shop

Figure 3.3: Rendered HTML of Code 3.1.

MCS, naive style and short text exclusion

Style	Characters	Characters discarded (less than 4 words)
h1\${ }	24	0
td\${ }	30	30
h3\${ }	20	20
h5\${ }	27	27
p\${ }	742	0
h6\${ }	40	14
b\${ }	11	11

Table 3.1: Naive styles and number of relevant characters.

Welcome to the Demo-Shop

[Navigation](#)[Products](#)[Sale](#)[About Us](#)

Terms and Conditions

1. Lorem Ipsum

dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

1.1 Donec quam

felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, **aliquet** nec, vulputate eget, arcu.

1.2 In enim justo, rhoncus

ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.

2. Aenean leo

ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.

Thanks for visiting Demo-Shop

Figure 3.4: Rendered HTML of Code 3.1.

Content Extraction – Processing Demo

Lowest Common Ancestor (threshold = 95%*)

Node (XPath)	Coverage	Depth
/html/body/div[2]	0.9664	1
/html/body/div[2]/p[1]	0.2355	2
/html/body/div[2]/p[2]	0.1984	2
/html/body/div[2]/p[3]	0.2529	2
/html/body/div[2]/p[4]	0.2796	2
/html/body/p	0.0336	1

Table 3.2: Nodes, coverage and depth.

Welcome to the Demo-Shop

Navigation Products Sale About Us

Terms and Conditions

1. Lorem Ipsum

dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

1.1 Donec quam

felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.

1.2 In enim justo, rhoncus

ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.

2. Aenean leo

ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.

Thanks for visiting Demo-Shop

Figure 3.5: Rendered Demo-Shop with colored *content* and *not-content* areas.

* threshold set to 95% for demonstration purposes due to the short document

1. Motivation

2. Research Question

3. Content Extraction

- Requirements
- Implementation
- Processing Demo

4. Hierarchy Extraction

- Requirements
- Implementation
- Processing Demo

5. Architecture & Additional Technology

6. Evaluation

- Content Extraction
- Hierarchy Extraction

7. Conclusion & Future Work

Hierarchy Extraction – Requirements

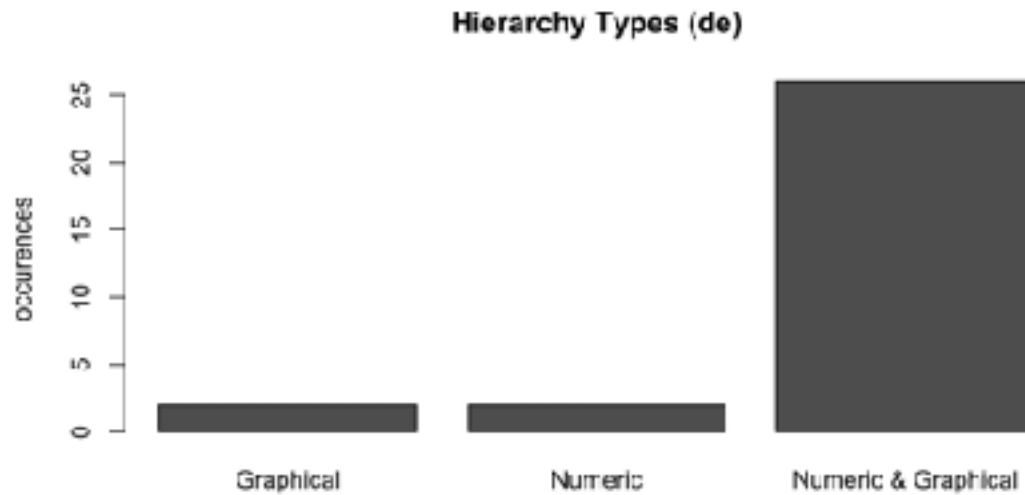


Figure 4.1: Hierarchy types in the German requirements sample.

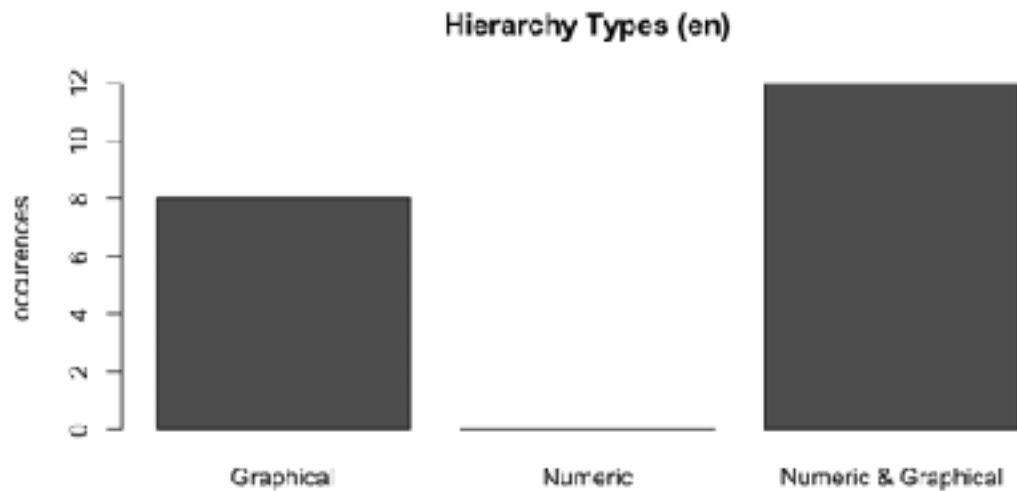


Figure 4.2: Hierarchy types in the English requirements sample.

Manual Review

Properties of the hierarchy representation:

- enumeration pattern (alphabetic, Arabic, Latin/Roman)
- headlines in a more prominent style, following paragraphs associated with the headline
- assumption: sections are not interrupted
 - exception: lists

Potential Problems:

- table of contents

[VLDB]Manabe, T., Tajima, K.: Extracting logical hierarchical structure of HTML documents based on headings.

[github]HEPS; [github]JTC-Detection-Corpus

Block Formation

Visual Based
Hierarchy Extraction

Enumeration Based
Hierarchy Extraction

Adjustments for
Lists ()

Rationale:

- headlines with more prominent styles preface their associated section(s)
 - assumption: sections are not interrupted
- enumeration patterns of different styles
 - Ignore enumeration patterns in the table of contents
- lists need special treatment
 - may violate the assumption of uninterrupted sections
 - each forms a section



1. group all text elements and include line breaks forced by HTML tags
2. create blocks from text sequences between line breaks
 - the block's style is determined by the majority of characters; characters in `<a>` are not considered
3. enumerations for each block are detected
 - `<li>` blocks receive a special numeration style (List)



Figure 4.3: Example of block formation, www.telescopehouse.com.



1. determine the MCS in the main content
 - rendered style and short text exclusion
2. identify the next headline style
 - headline styles are more prominent than the MCS
3. select all headlines of the current headline style and their associated content
 - associated content according to assumption: sections are not interrupted
4. continue within each of the content sections (from 2.)

Block Formation

Visual Based
Hierarchy Extraction

Enumeration Based
Hierarchy Extraction

Adjustments for
Lists ()

- enumerations are detected using a regular expression

- `\s[ \($]?([IVXLivx1]{1,7})|([0-9]{1,2})|[a-zA-Z])([ \. \- , : ]([IVXLivx1]{1,7})|([0-9]{1,2})|[a-zA-Z]))*[\- : \. )]? \s`

A: adjust hierarchy of existing visual based hierarchy using enumerations

1. determine the next enumeration style (i.e. Latin/Roman, alphabetic, Arabic)
2. validate numeration style
 - numeral: continuous numeration within enumeration pattern, at least 2 occurrences
 - table of contents: short blocks (less than 10 words) need to be followed by a content block
3. apply adjustment and continue in the subsections recursively

B: add hierarchy within existing non-headline content blocks using enumerations



- lists need special treatment
 - may violate the assumption of uninterrupted sections
1. check for list enumeration style within each content section
 2. adjust hierarchy according to list
 - content before and after the list on the same hierarchy level is associated with the same parent as the list content

Hierarchy Extraction – Processing Demo

Block Formation

ID	Text	Font-Size	Bold	Under-lined
1	Terms and Conditions	18px	✓	✗
2	1. Lorem Ipsum	13px	✓	✗
3	dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.	16px	✗	✗
4	1.1 Donec quam	10px	✓	✗
5	felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.	16px	✗	✗
6	1.2 In enim justo, rhoncus	10px	✓	✗
7	ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.	16px	✗	✗
8	2. Aenean leo	13px	✓	✗
9	ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.	16px	✗	✗

Table 4.1: List of all Blocks (incl. rendered style).

Welcome to the Demo-Shop

Irrelevant content

NavigationProductsSaleAbout Us

Terms and Conditions

1. Lorem Ipsum

dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

1.1 Donec quam

felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.

1.2 In enim justo, rhoncus

ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.

2. Aenean leo

ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.

Thanks for visiting Demo-Shop

Irrelevant content

Figure 4.4: Rendered Demo-Shop with colored *content* and *not-content* areas.

Hierarchy Extraction – Processing Demo

Visual Based

Determine MCS (rendered & short text exclusion):

Style	Characters	Characters discarded (less than 4 words)
18px, bold	20	20
13px, bold	27	27
16px	724	0
10px, bold	40	14

Table 4.2: MCS in relevant content.

Welcome to the Demo-Shop

Navigation Products Sale About Us

Terms and Conditions

1. Lorem Ipsum

dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

1.1 Donec quam

felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.

1.2 In enim justo, rhoncus

ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.

2. Aenean leo

ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.

Thanks for visiting Demo-Shop

Figure 4.4: Rendered Demo-Shop with colored *content* and *not-content* areas.

Hierarchy Extraction – Processing Demo

Visual Based I

ID	Text	Font-Size	Bold	Under-lined
1	Terms and Conditions	18px	✓	✗
2	1. Lorem Ipsum	13px	✓	✗
3	dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.	16px	✗	✗
4	1.1 Donec quam	10px	✓	✗
5	felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.	16px	✗	✗
6	1.2 In enim justo, rhoncus	10px	✓	✗
7	ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.	16px	✗	✗
8	2. Aenean leo	13px	✓	✗
9	ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.	16px	✗	✗

Table 4.3: Visual based hierarchy extraction, step 1.



blocks in current headline style

blocks in MCS



associated blocks (to preceding headline)

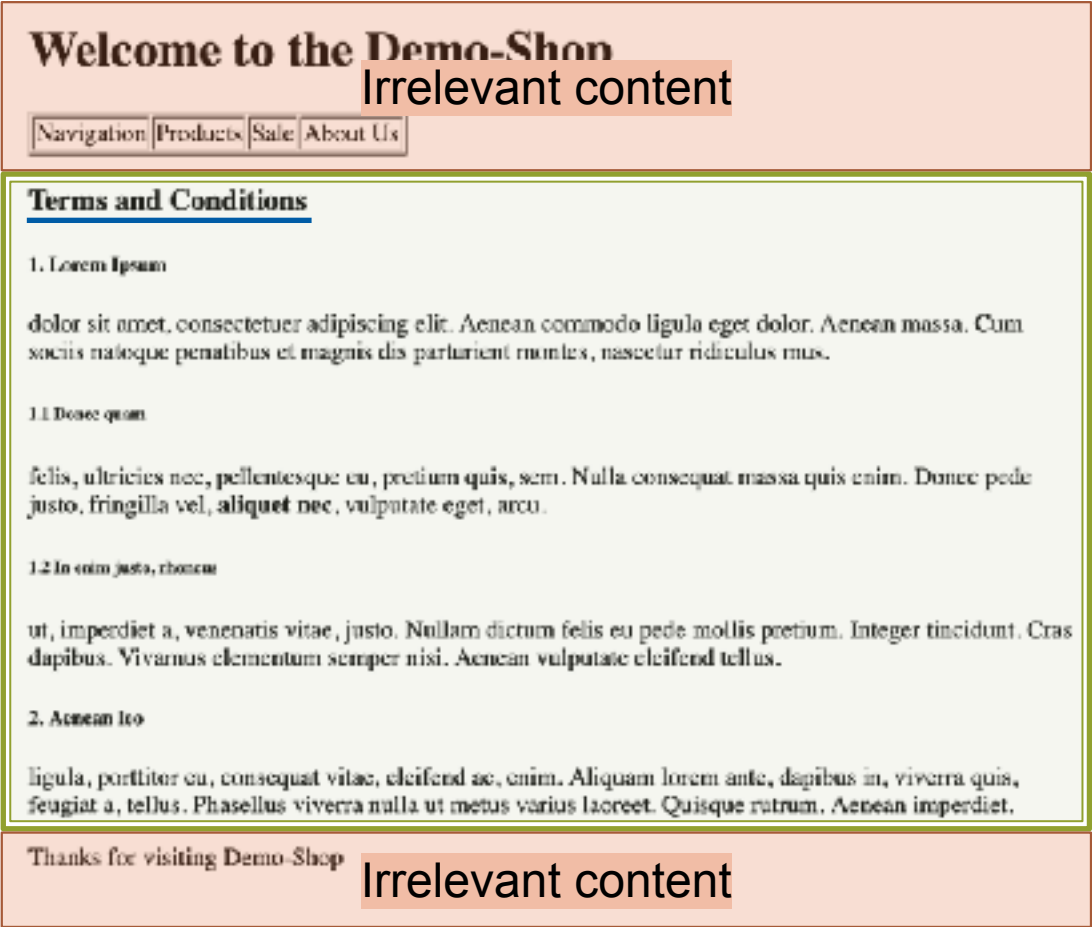


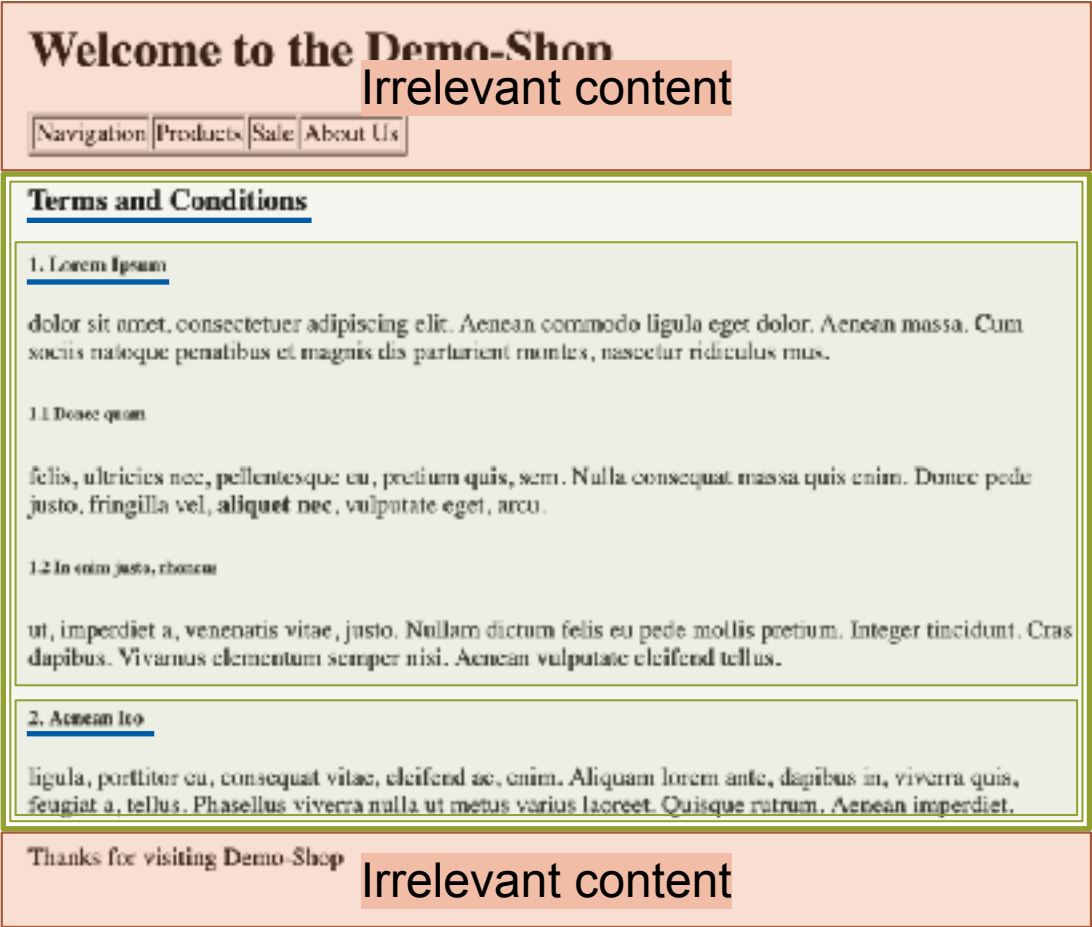
Figure 4.5: Rendered Demo-Shop with colored *content* and *not-content* areas and visual based hierarchy extraction, step 1.

Hierarchy Extraction – Processing Demo

Visual Based II

ID	Text	Font-Size	Bold	Under-lined
1	Terms and Conditions	18px	✓	✗
2	1. Lorem Ipsum	13px	✓	✗
3	dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.	16px	✗	✗
4	1.1 Donec quam	10px	✓	✗
5	felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.	16px	✗	✗
6	1.2 In enim justo, rhoncus	10px	✓	✗
7	ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.	16px	✗	✗
8	2. Aenean leo	13px	✓	✗
9	ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.	16px	✗	✗

Table 4.4: Visual based hierarchy extraction, step 2.



blocks in current headline style



blocks in MCS



associated blocks (to preceding headline)

Visual Based III

ID	Text	Font-Size	Bold	Under-lined
1	Terms and Conditions	18px	✓	✗
2	1. Lorem Ipsum	13px	✓	✗
3	dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.	16px	✗	✗
4	1.1 Donec quam	10px	✓	✗
5	felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.	16px	✗	✗
6	1.2 In enim justo, rhoncus	10px	✓	✗
7	ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.	16px	✗	✗
8	2. Aenean leo	13px	✓	✗
9	ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.	16px	✗	✗

Table 4.5: Visual based hierarchy extraction, step 3; done.



blocks in current headline style

blocks in MCS



associated blocks (to preceding headline)

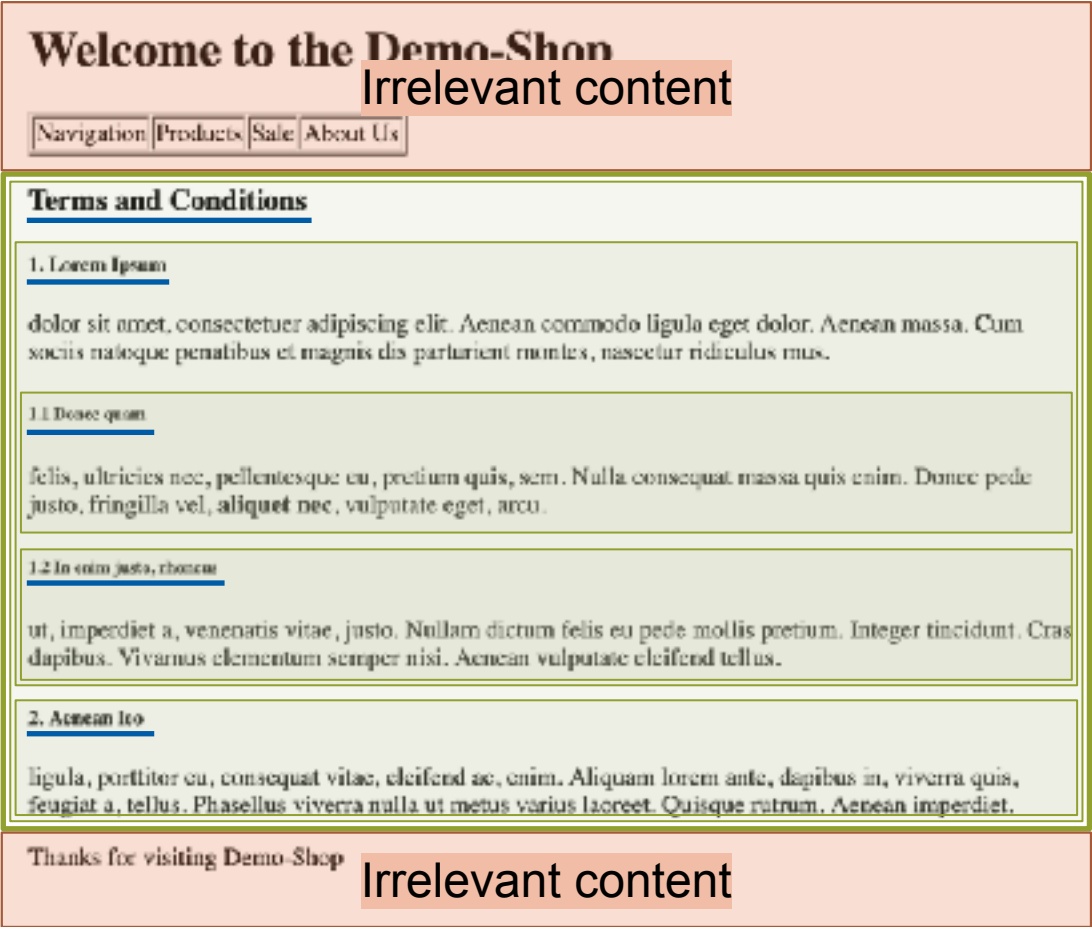


Figure 4.7: Rendered Demo-Shop with colored *content* and *not-content* areas and visual based hierarchy extraction, step 3; done.

Hierarchy Extraction – Processing Demo

Enumeration Based

ID	Enumeration	Pattern
1	[]	[]
2	[1]	[numeric]
3	[]	[]
4	[1, 1]	[numeric, numeric]
5	[]	[]
6	[1, 2]	[numeric, numeric]
7	[]	[]
8	[2]	[numeric]
9	[]	[]

Table 4.6: Detected enumerations & patterns.

Enumeration valid $\Rightarrow$ no adjustment needed

No `<li>` in html $\Rightarrow$ no adjustment needed

Welcome to the Demo-Shop

Navigation Products Sale About Us

Terms and Conditions

1. Lorem Ipsum

dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

1.1 Donec quam

felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.

1.2 In enim justo, rhoncus

ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus.

2. Aenean leo

ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet.

Thanks for visiting Demo-Shop

Figure 4.7: Rendered Demo-Shop with colored *content* and *not-content* areas, visual based hierarchy extraction and enumeration based hierarchy extraction incl. adjustments for lists.

1. Motivation

2. Research Question

3. Content Extraction

- Requirements
- Implementation
- Processing Demo

4. Hierarchy Extraction

- Requirements
- Implementation
- Processing Demo

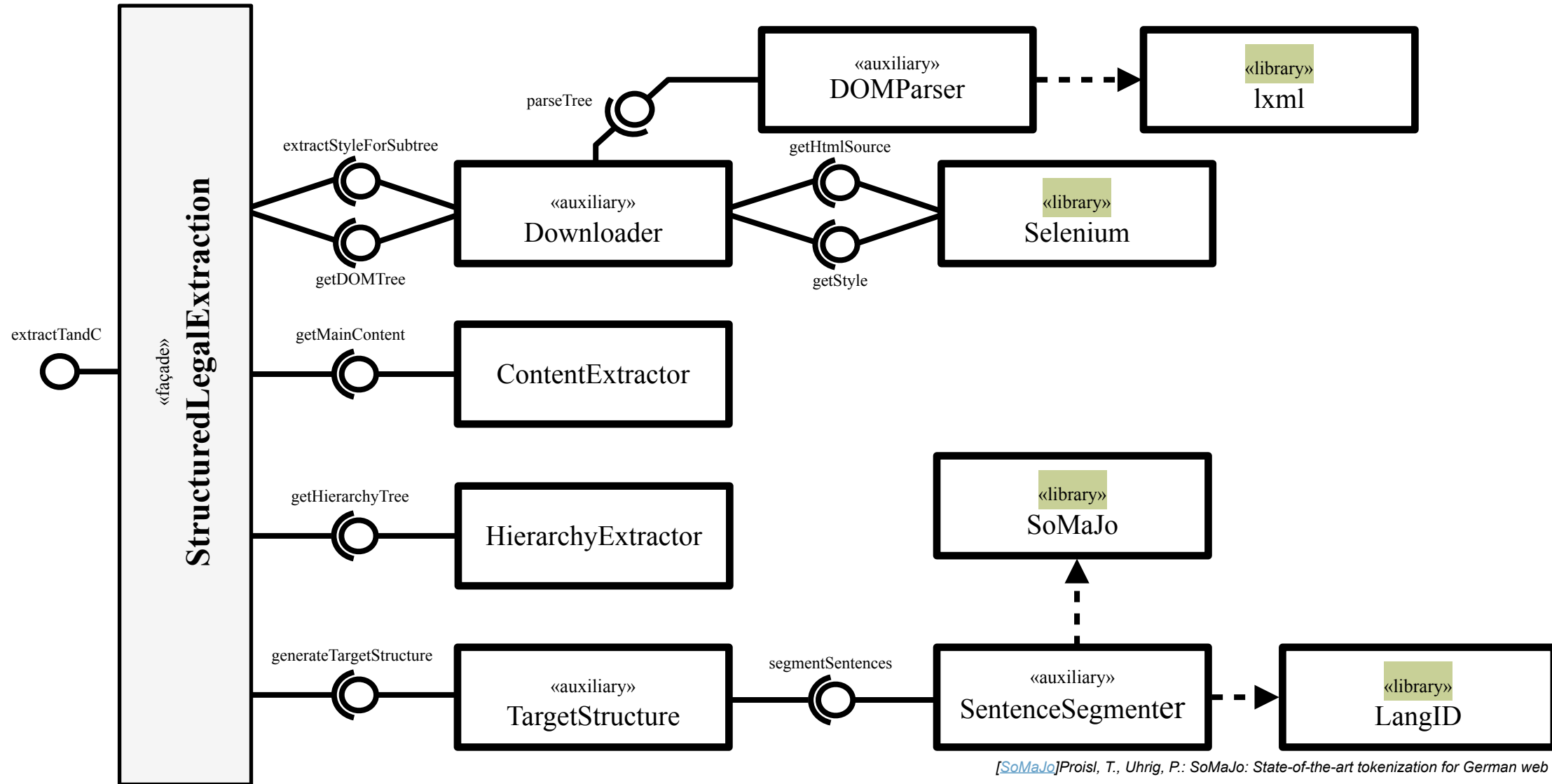
5. Architecture & Additional Technology

6. Evaluation

- Content Extraction
- Hierarchy Extraction

7. Conclusion & Future Work

Architecture & Additional Technology



[SoMaJo]Proisl, T., Uhrig, P.: SoMaJo: State-of-the-art tokenization for German web and social media texts

Figure 5.1: UML component diagram of the structuredlegalextraction library.

[github]SoMaJo; [github]Selenium; [github]LangID; [github]lxml

Architecture & Additional Technology

Web Page Download

- download full HTML
- extract CSS style
- full XPath support
- Java-Script support (*opt.*)

«library»
Selenium

HTML Parser

- full XPath support
- generate XPath from elements

«library»
lxml

Sentence Segmentation
and Tokenising

- applicable in the domain of legal documents

«library»
SoMaJo

Language
Determination

- support German and English language

«library»
LangID

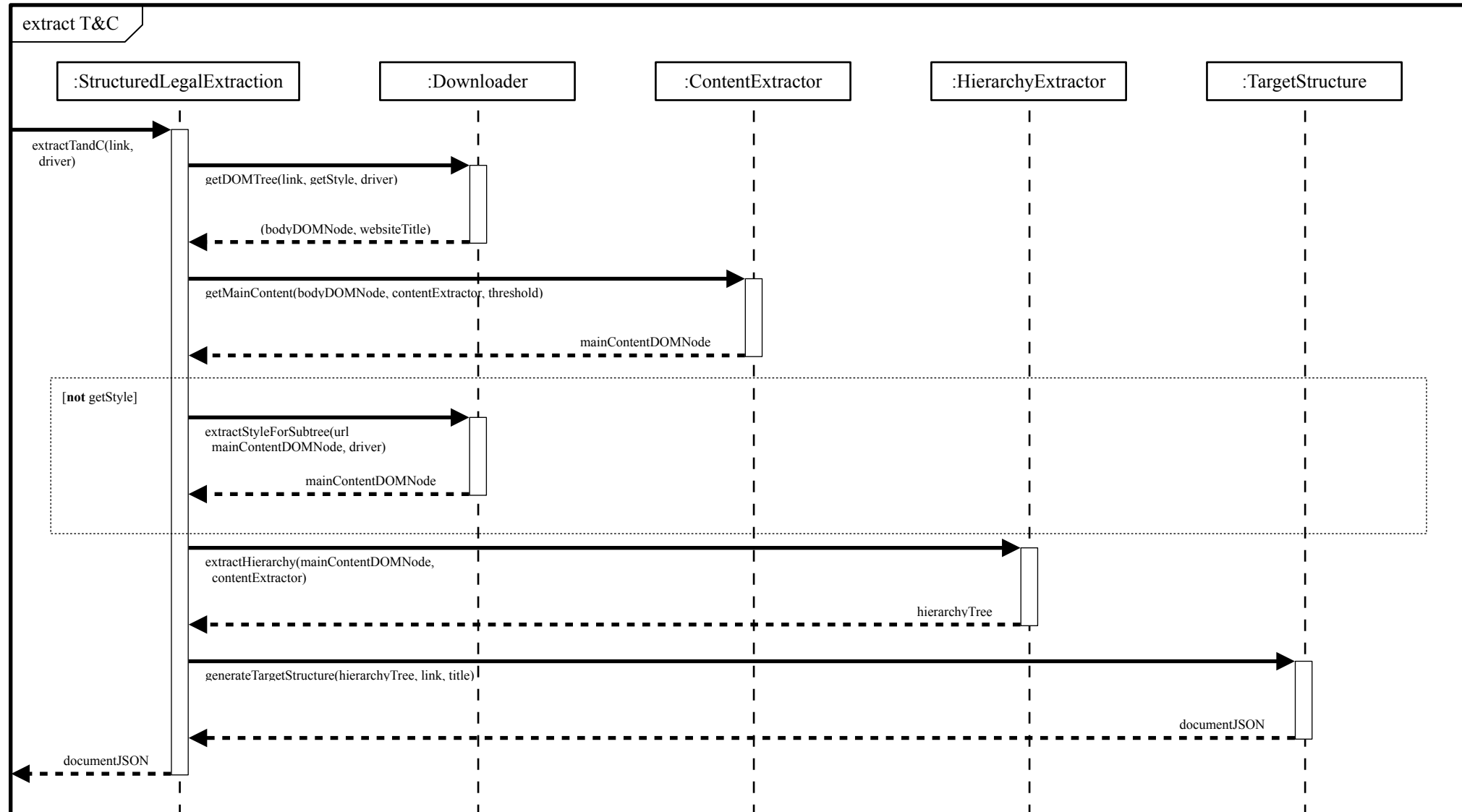


Figure 5.2: UML sequence diagram of the structuredlegalextraction libraries extractTandC method.

[[github](#)]SoMaJo; [[github](#)]Selenium; [[github](#)]LangID; [[github](#)]Ixml

1. Motivation

2. Research Question

3. Content Extraction

- Requirements
- Implementation
- Processing Demo

4. Hierarchy Extraction

- Requirements
- Implementation
- Processing Demo

5. Architecture & Additional Technology

6. Evaluation

- Content Extraction
- Hierarchy Extraction

7. Conclusion & Future Work

Evaluation – Content Extraction

Two samples used during evaluation:

- Requirements sample used to derive requirements (30 de, 20 en)*
- Test sample to validate results (30 de, 20 en)**

Extractor	too late	too early	correct
LowestCommonAncestor Extractor	1	3	45
Boilerpipe ArticleExtractor	16	4	24
Boilerpipe LargestContententExtractor	29	2	13
Boilerpipe CanolaExtractor	7	15	22
JusText	6	11	25
Trafilatura	5	2	37

Table 6.1: Start, requirements sample.

Extractor	too late	too early	correct
LowestCommonAncestor Extractor	3	2	44
Boilerpipe ArticleExtractor	23	13	8
Boilerpipe LargestContententExtractor	32	1	11
Boilerpipe CanolaExtractor	3	27	14
JusText	8	15	19
Trafilatura	5	4	35

Table 6.2: End, requirements sample.

Extractor	too late	too early	correct
LowestCommonAncestor Extractor - START	3	0	46
LowestCommonAncestor Extractor - END	0	2	47

Table 6.3: Start & End, test sample.

Performance in the center is not shown as *LowestCommonAncestor* Extractor extracts continuous content.

* *LowestCommonAncestor* Extractor processed 49 of 50 web pages, the other extractors processed 44 of 50 web pages from the requirements sample.

** *LowestCommonAncestor* Extractor processed 49 of 50 web pages from the test sample.

Evaluation – Content Extraction

Sources of errors:

1. Threshold too high $\Rightarrow$ no lowest common ancestor $\Rightarrow$ fallback algorithm is triggered
2. No container surrounding the main content $\Rightarrow$ no lowest common ancestor $\Rightarrow$ fallback algorithm is triggered
3. Different tags with the same style $\Rightarrow$ whenever one tag is held in its own container $\Rightarrow$ only part of the content extracted
 - Example: part of T&C in `<div>` followed by numerous `<li>` surrounded by `<ul>/<ol>`, see Figure 6.1

	TERMS AND CONDITIONS References in these Terms and Conditions to "you", "your", "user" and "customer" all relate to the party using the Website. References in these Terms and Conditions to Website refer to the Website http://www.escentual.com/ and references to "we", "us" and "Escentual" relate to RK Aggarwal Ltd T/A Escentual (Company Number 02031097) registered office: Churchgate House, Church Road, Whitthurch, Cardiff, CF14 2DX. We reserve the right, at our sole discretion, to change, modify, add, amended or remove any part of these Terms and Conditions and it is your responsibility as the consumer to check these Terms and Conditions from time to time. If you do not agree to our right to change the Terms and Conditions or the Website from time to time please do not use the Website. By accessing this Website you agree to these Terms and Conditions. If you do not agree to any of the stated Terms and Conditions you may not use or access this Website. The Website www.escentual.com is solely for personal use and the Website may not be used for any other uses. Please note that these Terms and Conditions do not affect your statutory rights as a consumer. 1. Shopping online 1.1. All products we sell are genuine and supplied by the brand owners, manufacturers or authorized UK distributors. 1.2. We take every care to ensure that the product images, description and specification of each product listed on the Website are accurate. However, although the colours in images on Website are a close representation a slight difference may occur in the actual product you receive. We always recommend trying the products that we sell in store before purchasing online. 1.3. By submitting an order to us through our Website, you confirm that you are the person referred to in your billing details and all of your details are valid and correct. 2. Acceptance of an order 2.1. You can place an order in one of three ways, either online, over the telephone through our customer service department or via post using Cheque, Postal Order or International Money Order.
--	--

Figure 6.1: Terms and Conditions of escentual.com, 210903

Evaluation – Hierarchy Extraction

Error scores:

- 0: each section with correct parent, correct content, and correct title
- 0.4: wrong parent
- 0.5: wrong content
- 0.1: wrong title

Error Share	Amount
0	12
(0; 0.05]	12
(0.05; 0.1]	1
(0.1; 0.15]	2
(0.15; 0.2]	1
(0.2; 0.3]	1
(0.3; 0.5]	1
(0.5; 1]	0
Failed*	0

Table 6.4: Error scores of hierarchy extraction for the German requirements sample.

Error Share	Amount
0	5
(0; 0.05]	4
(0.05; 0.1]	1
(0.1; 0.15]	2
(0.15; 0.2]	0
(0.2; 0.3]	0
(0.3; 0.5]	1
(0.5; 1]	0
Failed*	6

Table 6.5: Error scores of hierarchy extraction for the English requirements sample.

Error Share	Amount
0	8
(0; 0.05]	11
(0.05; 0.1]	1
(0.1; 0.15]	1
(0.15; 0.2]	1
(0.2; 0.3]	1
(0.3; 0.5]	5
(0.5; 1]	0
Failed*	2

Table 6.6: Error scores of hierarchy extraction for the German test sample.**

Error Share	Amount
0	9
(0; 0.05]	3
(0.05; 0.1]	2
(0.1; 0.15]	1
(0.15; 0.2]	2
(0.2; 0.3]	2
(0.3; 0.5]	1
(0.5; 1]	0
Failed*	0

Table 6.7: Error scores of hierarchy extraction for the English test sample.

$$x \in [a; b) \mid x \geq a \wedge x < b \quad \text{error share} = \frac{\text{sum of total error scores}}{\text{sum of total nodes extracted}}$$

* Failed content extraction.

** One page was no longer available during the evaluation of the hierarchy extraction.

Sources of errors:

1. Use of bold/large text, i.e. violation of assumptions about headline styles $\Rightarrow$ regular content is identified as headline (see Figure 6.2)
2. Wrong enumeration $\Rightarrow$ sequence of numeration is not valid $\Rightarrow$ no hierarchy detected
3. Interrupted sections, i.e. violation of the assumption "Sections are not interrupted" $\Rightarrow$ text is assigned to the wrong headline
4. Use of tables $\Rightarrow$ each cell is its own block $\Rightarrow$ many blocks in different styles/many numbers in the table $\Rightarrow$ wrong visual/enumeration based adjustments
5. Failed style extraction $\Rightarrow$ standard style is set $\Rightarrow$ wrong visual based separation

regular text highlighted →	9. Speicherung des Vertragstextes Wir speichern die Bestelldaten und den Vertragstext. Bei Abschluss der Bestellung ist der Vertragstext aus Sicherheitsgründen nicht mehr über das Internet zugänglich. Soweit Sie sich als Kunde registrieren haben, können Sie den Vertragstext über einen Login mindestens bis zur Erfüllung des Vertrages einsehen. Durch die Druckfunktion Ihres Browsers haben Sie die Möglichkeit, den Text selbst zu speichern. Die AGB können Sie jederzeit auf dieser Seite einsehen. Nach Abschluss der Bestellung werden Ihnen die Bestelldaten und die AGB einschließlich der Widerrufsbelehrung und des Widerrufsformulars per Email übersandt.
	10. Gewährleistungsgarantie Die Gewährleistung erfolgt nach den gesetzlichen Bestimmungen.
	Wir geben keine eigenen Garantien für unsere Produkte. Regelmäßig gewähren die Hersteller Garantien auf ihre Produkte, die meist 24 Monate gültig sind. Die Details hierzu entnehmen Sie den Garantiebedingungen des Herstellers, die der Warelieferung beigelegt sind. Ihre Gewährleistungsrechte gegenüber uns werden hierdurch nicht eingeschränkt. Sie können uns daher bei Mängeln der Sache unabhängig von der Inanspruchnahme der Herstellergarantie im Rahmen der gesetzlichen Gewährleistungsbestimmungen jederzeit in Anspruch nehmen. Bei gebrauchten Waren beträgt die Gewährleistungsfrist ein Jahr ab Lieferung der Ware. Die verkürzte Gewährleistungsfrist gilt nicht für Schäden aus der Verletzung des Lebens, des Körpers oder der Gesundheit, die vor uns schuldhaft verursacht wurden oder für sonstige Schäden, die auf einer vorsätzlichen oder grob fahrlässigen Pflichtverletzung beruhen.
	11. Haftung Wir haften uneingeschränkt für Schäden aus der Verletzung des Lebens, des Körpers und/oder der Gesundheit. Ferner haften wir im Fall des Vorsatzes oder grober Fahrlässigkeit bei arglistigem Verschweigen eines Mangels oder in allen anderen gesetzlich geregelten Fällen.

Figure 6.2: Terms and Conditions of steber.de, 210903

1. Motivation

2. Research Question

3. Content Extraction

- Requirements
- Implementation
- Processing Demo

4. Hierarchy Extraction

- Requirements
- Implementation
- Processing Demo

5. Architecture & Additional Technology

6. Evaluation

- Content Extraction
- Hierarchy Extraction

7. Conclusion & Future Work

RQ1: What are special requirements for the extraction of contracts in comparison with *regular* (i.e. news articles, blogposts) content?

- continuous content
- addresses and phone numbers are part of the relevant content (no hint for imprint/footer)

RQ2: How to extract the relevant parts from contracts in raw HTML?

- *LowestCommonAncestor Extractor*

RQ3: How to extract the structure and the hierarchy of paragraphs, (sub-)titles, and related clauses?

- rule based approach ignoring structural information except for and line breaks (blocks)
- enumeration patterns and visual styles

Future Work:

- evaluate *LowestCommonAncestor Extractor* in other domains
- address the sources of error in hierarchy extraction and improve time efficiency of rendered style extraction



Tobias Schamel

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

www.matthes.in.tum.de

tobias.schamel@tum.de

