

# Automatische Sprachübersetzung von LATEX-Dokumenten

Name: Hendrik Theede

Matrikelnummer: 221201256 Abgabedatum: 02.12.2025

Betreuer und Gutachter: Prof. Dr. rer. nat. habil. Clemens H. Cap

Universität Rostock

Fakultät für Elektrotechnik und Informatik

Bachelorarbeit am Lehrstuhl für Informations- und Kommunikationsdienste Fakultät für Elektrotechnik und Informatik

# Abstrakt

placeholder

## Inhaltsverzeichnis

1	Einl	eitung 1
	1.1	Hintergrund
	1.2	Anforderungen
2	Prol	blemfälle 3
	2.1	Translativ
		2.1.1 Sonderzeichen
		2.1.2 Leerzeichen
		2.1.3 Zeilenbrüche
		2.1.4 Dokumentenbrüche
	2.2	Technisch
		2.2.1 Hilfsdateien
		2.2.2 Unerreichbare Informationen
		2.2.3 Figuren und Tabellen
		2.2.4 Literaturverzeichnisse
		2.2.5 Category Codes
	2.3	Spezifischer Technologien
		2.3.1 Kommentare
		2.3.2 Dilemmatische Makros
		2.3.3 TikZ und Layouting
		2.3.4 Quellmehrsprachigkeit
	2.4	Sprachliche Schwierigkeiten
		2.4.1 Glossare und Nomenklaturen
		2.4.2 Weitere
3		hnologien 13
	3.1	Übersicht
		3.1.1 Auflistung
		3.1.2 Eingrenzung
		3.1.3 Auswertung
	3.2	Einschätzung
	3.3	Fazit
4	Offe	ene Problematiken 14
	4.1	Verfolgte Ideen
	4.2	Gelöste Probleme
	4.3	Lessons Learned
	7.J	Ecosons Ecarricu

5	Fazit			
	5.1 Zusammenfassung	15		
	5.2 Ausblick	15		
	5.3 Weiteres	15		
6	6 Eigenständigkeitserklärung			
Lit	eratur	17		
	Anhänge A.1 Fontskalierung auf Webseiten	<b>18</b>		

### 1 Einleitung

#### 1.1 Hintergrund

Der erste Satz dieses Werkes befindet sich noch in Arbeit. Vermutlich wird dieser mit einem Wort beginnen, welches seinerseits den Anfangsbuchstaben "a" trägt (bsps. "Ausgehend (von)" oder "Anders (als . . . behaupten)").

Herkömmliche Software zur Übersetzung von menschlicher Sprache auf TFX-Quellcode anzuwenden, erzeugt schnell Dokumente, welche entweder nicht vollständig übersetzt wurden oder sich nicht mehr kompilieren lassen. Mit Hilfe von Google Translate lassen sich wesentliche Gründe hierfür finden und wie sich diese äußern. Beispielsweise führt eine Übersetzung von hello wor\textit{ld} nicht zu Hallo We\textit{lt}, sondern zu hallo wor\textit{ld}. Abgesehen von der Frage, wo die kursive Hervorhebung im eigentlichen String erfolgen soll, werden Leser eines kompilierten Dokumentes das Wort "Welt" erkennen. Zuvor beschriebene Zeichenkette wird von TFX zu "hello world" aufgelöst, in welcher das Wort "world" für einen menschlichen Leser als das englische Wort für "Welt" erkenntlich bleibt. Fehlt die Kenntnis über eine der Sprachen (DE,EN), würde einem monolingualen Leser Teil der Wortkette geraubt werden. Selbstverständlich sind die Wörter "world" und "Welt" einander sehr nahe und auch eine Formulierung der Art "Hallo Welt" lässt Vermutungen gegenüber eines größeren Kontexts zu. Anders wäre dies, wenn das Auslassen von auch nur einem Wort keine Rückschlüsse mehr auf einen größeren Kontext mehr zulässt. \$\mathbb{P}\\$robability density function wäre ein denkbarer stilistischer Weg bereits in z.B. einem Folientitel bereits eine Notation für eine Wahrscheinlichkeitsdichtefunktion einzuführen. Hierbei würde der Verlust des Wortes "probability" den stochastischen Kontext aufheben. Der Verlust des Wortes "density" würde einen Kontext innerhalb der Stochastik verändern und ohne das Wort "function" ist fraglich, wovon die Rede ist. Vor allem in größeren Dokumenten könnten hierdurch Logikbrüche entstehen.

Eine Betrachtung eines "Übersetzers" als Konzept veranschaulicht die Problematik auf abstrakterer Ebene. Sollte der Kontext des Dokumentes unbekannt sein, werden sich unausweichlich semantische Fehler einschleichen. Bereits das gezeigte Beispiel könnte für z.B. eine Folie einer Lesung den restlichen Kontext der Seite entfernen und dadurch die Möglichkeit bieten umgangssprachliche Bedeutungen in Wörter zu interpretieren, anstatt einer Mathematischen (bspw. "ungerade" könnte im Englischen "crooked", statt "odd" produzieren). Noch weitere sprachliche Beispiele finden sich schnellig durch Wörter mit zeitlichem/räumlichen Bezug. Der Satz Morgen wird es regnen. könnte ohne das Wort "morgen" als Frage mit unzureichend eingehaltener deutscher Grammatik interpretiert werden. (*Wird es regnen?*). Hierbei verliert man eine getroffene Aussage über das Wetter, welches bekanntlicherweise nur schwer vorhergesagt werden kann.

#### 1.2 Anforderungen

Genauso wie das Fehlen einzelner Wörter die sprachliche Bedeutung für einen Menschen brechen kann, treten ähnliche Probleme auch in TEX auf. Einzelne LATEX Makros nicht zu übersetzen ist unbedeutend, da sie ihre Bedeutung für einen TEX Compiler behalten. Alleine einzelne Wörter eines Makros zu übersetzen kann dazu führen, dass größere Inhalte (im Sinne: Menge an Worten) nicht mehr in einem kompilierten Dokument vorzufinden sind, was auf eine Fähigkeit von TEX zurückführbar ist. Die Möglichkeit in bestimmten Fällen eine Dateiendung auszulassen, führt beim Einbinden von anderen .tex Dateien in einem TEX Dokument zu fehlerhaften/fehlenden Ressourcenangaben. \include{clock} zu \include{Uhr} zu übersetzen (wie bspw. Google Translate am 06.10.2025) würde nun nicht mehr zu \include{clock.tex} aufgelöst werden, sondern zu \include{Uhr.tex} (bei welchem nicht davon auszugehen ist, dass diese Datei zur Kompilierzeit im System zwingend vorliegt).

Daher muss nach einer Lösung gesucht werden, welche diese technischen und sprachlichen Hürden überwinden kann. Neben solchen rein technischen Details, darf die Perspektive des Lesers (wörtlich) nicht missachtet bleiben und keine Übersetzungsprozesse dürfen zu versteckten Inhalten im Dokument führen. Diese Verbergung resultiert aus verschiedenensten Layouting-Problemen, ähnlich wie bei der Skalierung von Boxen auf Webseiten (Anhang A.1) und ist abhängig von einzelnen Sprachen dazu in der Lage unbemerkt verdeckte textliche Inhalte zu provozieren. Wünschenswert ist neben vorigen Aspekten auch Möglichkeiten für den Endnutzer zu erlauben, sollte dieser spezielle Übersetzungen oder Kontexte für einige Wörter wünschen, welche jedoch nicht aus dem Dokument selbst hervorgehen. Außerdem sollte ein möglichst hoher Support für sowohl verschiedene menschliche Sprachen, aber auch verschiedene LATEX-Pakete gegeben sein, wobei Letzteres nur ein Bonus ist, sollten Systeme wie TikZ, bzw. pgfplots oder BibTEX innerhalb LATEX (zusammen mit TEX) nutzbar bleiben.

<sup>&</sup>lt;sup>1</sup>Zwei adjazente Textfelder müssen sich zwangsläufig überlagern, wenn eines unabdingbar größer werden muss, da z.B. die Textgröße nicht verkleinert werden kann und das Wachstum eines Textfeldes nur in das Gebiet eines Anderen stattfinden kann. Dies wäre z.B. mit Hilfe von Inhaltsangaben auf Lebensmitteln vorstellbar. Sollten diese Wörter übersetzt werden und dadurch mehr Textfläche nach rechts benötigen, würden sie in den tabellarischen Bereich der eigentlichen quantitativen Angaben des z.B. Brennwertes, der Makro- sowie der Mikronährstoffe, hereinragen, wodurch das Risiko besteht, dass diese verdeckt werden.

#### 2 Problemfälle

Uneindeutigkeiten in der Sprache sind für einen Leser oft schwer nachzuvollziehen. In TEX muss allerdings zu mindestens einem Zeitpunkt die Information über das Aussehen des entgültigen Dokumentes in einer modellartigen Form vorliegen. Diese Information kann in LATEX verborgen sein oder aber bei einem Übersetzen verloren gehen. Die Probleme unterteilen sich in verschiedene Fälle und werden hinsichtlich des Kontextes dieses Informationsverlustes in translative (beim Übersetzen), technische (LATEX), spezifische technische (in Kombination mit TEX nutzbare Programme) und sprachliche Probleme. Sprachliche Probleme verursachen teilweise dillematische Probleme,welche als "Schwierigkeiten" und nicht als zu lösende Probleme dargestellt werden. Einzelne aufgeführte Beispiele zur Veranschaulichung beschriebener Probleme sind mit Hilfe von nicht spezifischer Software erzeugt (Google Translate), um zu zeigen, dass jeweilige Situation in einer Software zur Übersetzung von menschensprachlichen Inhalten entstehen könnten.

#### 2.1 Translativ

#### 2.1.1 Sonderzeichen

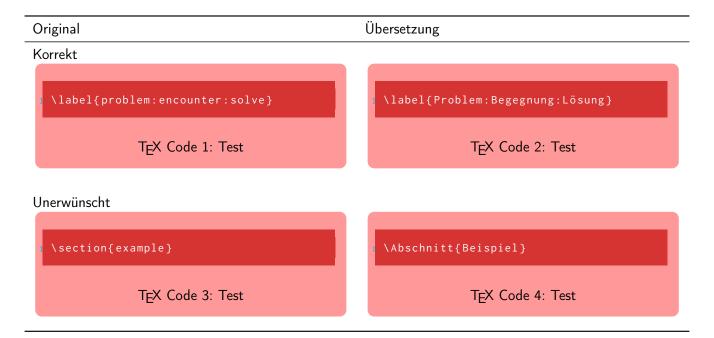


Tabelle 1: Fehler in einem Token

Beschreibung und Begründung Der für einen TEX Compiler relevante Befehl \label bleibt unverändert, allerdings section wird fälschlicherweise als Abschnitt übersetzt. Warum label nicht erfasst werden sollte, wenn die folgenden drei Wörter übersetzt wurden, wirft Fragen auf. Zu sehen ist ein String, welcher menschliche

Sprache mit Sonderzeichen vermischt. Da insbesondere Klammern in (vielen) sprachlichen Kontexten hilfreich sind, werden deren Inhalte selbst zunächst nach zusammenhängenden Worten durchsucht, welche ihrerseits durch : getrennt sind, oder als Klammern betrachtet werden können. Ein Entfernen der Klammern lässt lässt in erstem Beispiel \label problem encounter solve und in zweitem Beispiel \section example stehen. Zu sehen ist hier also bereits, dass Google Translate bei einer, wenn man es so interpretieren möchte, Vernestung zweiten Grades scheitert, jedoch einfache Vernestungen noch erkennt<sup>2</sup>.

**Takeaway** Teile der T<sub>E</sub>X-Syntax lassen sich anhand von \, {, }, [, ], \$, \$\$ oder \% erkennen und müssten daher ausgeschlossen werden. Anders als in mathematischen Formeln zeigen sich Sonderzeichen jedoch nicht paarweise auf, sodass sie nicht paarweise ignoriert werden können. Man kann sich diese Art von Fehlern wie 0-dimensionale Fehler vorstellen, wobei die nullte Dimension hierbei bei einem einzelnen Wort beginnt (welche als Punkte zu verstehen sind).

#### 2.1.2 Leerzeichen

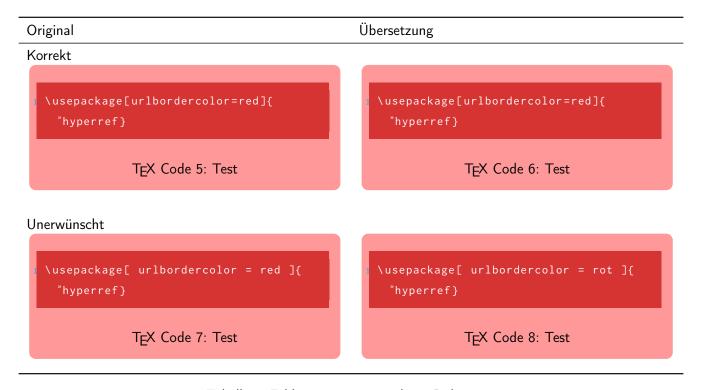


Tabelle 2: Fehler in einem einzeiligen Dokument

<sup>&</sup>lt;sup>2</sup> Vernestung: Klammern in Klammern, wie in der Mathematik

Beschreibung und Begründung Die Optionen innerhalb eckiger Klammern lassen auch Whitespace zu. Dies kann jedoch für die Nutzung einiger Funktionen in z.B. wichtigen Paketen wie hyperref dazu führen, dass falsche Wörter übersetzt werden, die ein Kompilieren des Dokumentes verhindern.

Abstrahierung Teile der TEX-Syntax lassen sich nicht nur anhand der zuvor beschriebenen Zeichenketten erkennen, sondern lassen sich auch in Zeilen wiederfinden. Diese Art von Fehlern bahnt den Weg zu einer Dimension, wodurch nicht nur innerhalb eines Wortes (Punktes), sondern auch zwischen verschiedenen Punkten Fehler entstehen könnten (also innerhalb einer Zeile).

#### 2.1.3 Zeilenbrüche

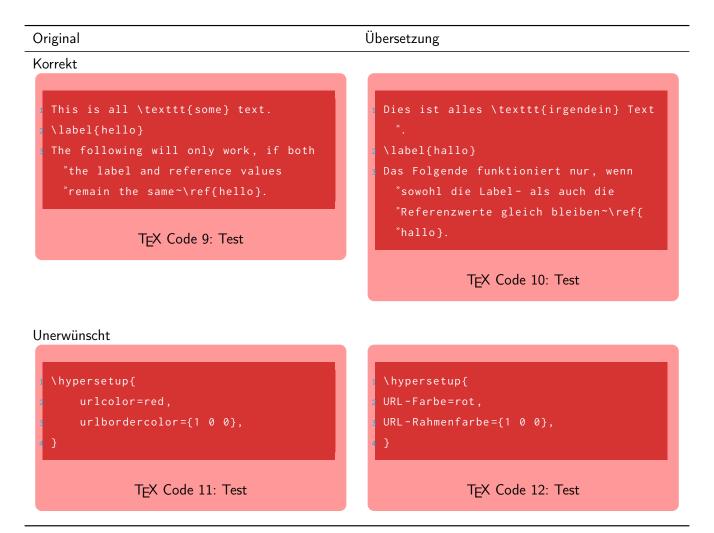


Tabelle 3: Fehler in einem einzeiligen Dokument

#### Beschreibung

**Abstrahierung** Teile der TEX-Syntax lassen sich nicht nur anhand von einzelnen Zeilen oder Zeichenketten erkennen, sondern könnten sich auch in verschiedenen Zeilen wiederfinden lassen. Diese Art von Fehlern kann 2-dimensional betrachtet werden, wodurch Fehler auch zwischen Zeilen entstehen können.

#### 2.1.4 Dokumentenbrüche

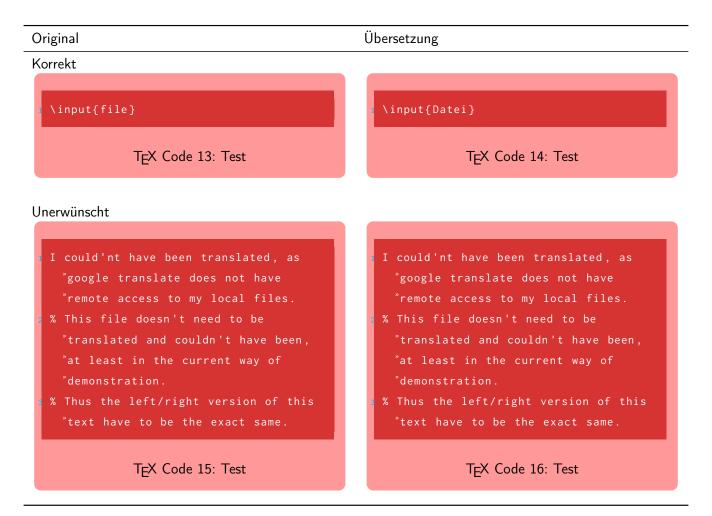


Tabelle 4: Fehler in einem einzeiligen Dokument

#### Beispiel

**Beschreibung** Die Übersetzung von Datei x, welche Datei y via input oder include führt zwar dazu, dass Datei x übersetzt wird, aber Datei y nicht.

Abstrahierung Teile der TEX-Syntax müssen nicht zwingend in einer Datei vorliegen, sondern könnten auch in verschiedenen Dateien integriert sein. Die Klassifizierung simpler Probleme gelangt in TEX hier bereits in der dritten Dimension an, weswegen sich fortan bereits mit "fortgeschrittenen" Problemen beschäftigt werden muss (welche sich Teils über mehrere "Dimensionen" erstrecken). Eine vierte Dimension existiert physikalisch nicht,

ist jedoch mathematisch formulierbar<sup>3</sup> und äußert sich in diesem Kontext auf eine Erhöhung von Laufzeitkomplexitäten.

#### 2.2 Technisch

#### 2.2.1 Hilfsdateien

2.2.1.1 Struktur dieses Abschnittes Eine Datei nutzt ein Literaturverzeichnis (BibTEX), pdfTEX (\pdfcomment{}), \footnote, ein Inhaltsverzeichnis, ein Glossar,

#### 2.2.1.2 Beispielliste

Inhaltsverzeichnisse, Abbildungsverzeichnisse, Tabellenlisten

#### Beispiel

Erläuterung Ändern wir den Titel eines Paragraphen oder Abschnittes, dann erfasst dies der TEX Compiler beim ersten Durchlauf. Jedoch der String im Inhaltsverzeichnis kann nur verändert werden, sobald diese Information zu Beginn des nächsten Kompilierungsprozesses in der entsprechenden Hilfsdatei vorliegt (.toc).

#### 2.2.1.2.1 Backrefs

**Beispiel** Bedarf evtl. einer bildlichen Veranschaulichung. Meint: Inhaltsverzeichnis, Tabelle,...vor einer backwards reference verschiebt die echte Position der (Phantom-) Sektion. Daher muss zunächst bestimmt werden, wo im Dokument eine Referenz auf ein existierendes Label stattfindet, welches vorherig bereits vergeben wurde

**Erläuterung** Ein Verweis auf einen vorherigen Paragraphen kann nur klickbar verlinkt werden, wenn die Information, an welcher Stelle er sich im Dokument befindet, bereits klar ist. Da nach der Referenz weitere Abschnitte folgen können, welche vorherige Elemente mit variabler Größe verändern könnten, muss zunächst die Größe dieser bestimmt sein und gegenüber dieser kann dann der

#### Literaturverzeichnisse

#### **Beispiel**

#### Erläuterung

<sup>&</sup>lt;sup>3</sup>physikalisch: die vierte Dimension ist die Zeit, wenn man eine nicht-euklidische 3-dimensionale Bewegung verlangt (Teleportation)

PDF Funktionen			
Beispiel			
Erläuterung			
"footnotes"			
Beispiel			
Erläuterung			
2.2.1.3 Abstrahiertes Problem Falls auf Hilfsdateien zugegriffen wird, ist das mehrfache Kompilieren eines Laufzelt Dokumentes unvermeidbar. Gegeben $n_h$ womoglich existierenden Hilfsdateien, welche alle ihrerseits zu übersetzende Inhalte beinhalten können, folgt eine minimale Laufzeit $\mathcal{O}(n_h)$ .			
2.2.2 Unerreichbare Informationen			
Beispiele			
<b>Beschreibungen</b> Ein Dokument erwähnt ein Werk, in welchem es um die C-Programmierung geht. Rein aus den im System vorliegenden Dateien ist kein Kontext für das Wort "String" erkennbar, sodass ein Zugriff auf eine externe Ressource unabdingbar ist.			
<b>Abstrahierung</b> Einfache Cloud-Architektur. Ein Client möchte auf ein beliebiges Wissen einer Webseite (bzw. dem Server und den beanspruchten Speicherplätzen in einem (beliebigen) Rechenzentrum <sup>4</sup> zugreifen).			
2.2.3 Figuren und Tabellen			

Beispiele

Beschreibungen

Abstrahierung

#### 2.2.4 Literaturverzeichnisse

Beispiele

<sup>&</sup>lt;sup>4</sup>Hierbei ist nicht von Festspeicher zu reden. Aus Sicherheitsgründen sei davon auszugehen, dass sich die physischen Adressen des wissensrepräsentierenden Speichers regelmäßig und unvorhersehbar ändern

Original Übersetzung

Dokument

```
1 \documentclass{standalone}
2 \usepackage{natbib}
3 \bibliographystyle{agsm}
4 \begin{document}
5 ~\cite{salomon_c} %Kapitel 5
6 refers to character arrays as a string.
7 \bibliography{example_original.bib}
8 \end{document}
8 \end{document}
8 \end{document}
8 \end{document}
1 \documentclass{standalone}
2 \usepackage{natbib}
8 \bibliographystyle{agsm}
4 \begin{document}
5 ~\cite{salomon_c} %Kapitel 5
6 bezeichnet Zeichenarrays als Zeichenfolge.
7 \bibliography{example_original.bib}
8 \end{document}
8 \end{document}
```

#### Bibliothek

Tabelle 5: Beispiel für einen verpassten literarischen Kontext

**Beschreibungen** BibTEX erlaubt es an vielerlei Stelle eigene Strings in einer kompilierten TEX-Datei zu verbergen.

#### **Abstrahierung**

#### 2.2.5 Category Codes

Beispiele

Beschreibungen

**Abstrahierung** 

#### 2.3 Spezifischer Technologien

Mitunter NP-schwer.

#### 2.3.1 Kommentare

Beispiele

**Beschreibungen** Wohingegen sich 2.3.1 nicht mit anderen, in Kommentaren referenzierten, Dateien beschäftigt, soll sich hier auf solche Fälle konzentriert werde.

**Abstrahierung** Hier treffen simple Fehler aus den ersten drei Kategorien (in 2.1.1, 2.1.2 und 2.1.3 geschildert) aufeinander. In die dritte Dimension, also in andere Dateien, wird jedoch (vorerst) nicht traversiert, da auskommentierte Datei-Einbindungen nicht erfasst werden dürften. Ausgehend von ?? wird nun erwartet, dass eine Referenzierung von Dateien erwartet wird, welche sich in Kommentaren verbergen. Dies kann jedoch 2.3.4 beinhalten.

#### 2.3.2 Dilemmatische Makros

Beispiele

Beschreibungen

**Abstrahierung** 

#### 2.3.3 TikZ und Layouting

Beispiele

Beschreibungen

**Abstrahierung** 

#### 2.3.4 Quellmehrsprachigkeit

Beispiele

Beschreibungen

**Abstrahierung** Quelltexte anderer Quellsprachen (Programmiersprachen) können ihrerseits auf andere Dateien verweisen, oder andere Syntaktik tragen. Das Erkennen dieser ist theoretisch gesehen leicht, jedoch praktisch gesehen schnellig zu übersehen.

## 2.4 Sprachliche Schwierigkeiten

#### 2.4.1 Glossare und Nomenklaturen

Beispiele

Beschreibungen

Abstrahierung

2.4.2 Weitere

Beispiele

Beschreibungen

Abstrahierung

# 3 Technologien

- 3.1 Übersicht
- 3.1.1 Auflistung
- 3.1.2 Eingrenzung
- 3.1.3 Auswertung
- 3.2 Einschätzung
- 3.3 Fazit

## 4 Offene Problematiken

- 4.1 Verfolgte Ideen
- 4.2 Gelöste Probleme
- 4.3 Lessons Learned
- 4.4 Fazit

- 5 Fazit
- 5.1 Zusammenfassung
- 5.2 Ausblick
- 5.3 Weiteres

6	<b>Eigenst</b>	ändia	<b>kaitsar</b>	klärung
U	Ligensi	anuig	VEITZEI	Kiai ulig

J	Ligenstandigkertserklarung
habe benu	versichere hiermit, dass ich die vorliegende Arbeit selbstständig angefertigt und ohne fremde Hilfe verfasst zu. Dazu habe ich keine außer den von mir angegebenen Hilfsmitteln und Quellen verwendet und die den utzten Werken inhaltlich und wörtlich entnommenen Stellen habe ich als solche kenntlich gemacht. Ich chere, dass die eingereichte elektronische Fassung mit den gedruckten Exemplaren übereinstimmt.
Rost	ock, den 02.12.2025
Hend	drik Theede

# Literatur

## A Anhänge

#### A.1 Fontskalierung auf Webseiten

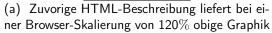
Beispielsweise produziert die folgende HTML-Notation bei einer Skalierung im Browser von 120 Prozent (Abbildung 2a) und 50 Prozent (Abbildung 2b) jeweilig zwei verschiedene PDF (unter welchen nur Zweitere alle textlichen Inhalte offenbart). Ähnliches kann auch innerhalb T<sub>F</sub>X geschehen, sollte

```
<html>
    <head>
        <title>Example</title>
        <style>
            /*formatting options are: none and black*/
                font-size:13em;
                height:50%;
            }
            /*formatting option: none = no background, black, courier*/
                font-family: 'Courier New', Courier, monospace;
            /*formatting option: black = black background, white, serif*/
            .t#black{
                background-color:black;
                color:white;
                margin-top: -2em;
        </style>
    </head>
    <body>
        <div class="t" id="none">Test</div>
        <div class="t" id="black">Test2</div>
    </body>
</html>
```

Abbildung 1: HTML-Beschreibung einer Webseite mit zwei Textflächen

Abbildung 2: Um die Dokumente von der restlichen Papierfläche abzugrenzen wurden schwarze Rahmen mittels TikZ hinzugefügt.







(b) Zuvorige HTML-Beschreibung liefert bei einer Browser-Skalierung von 50% obige Graphik