

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	Einle	inleitung 1				
	1.1	Hintergrund	1			
	1.2	Anforderungen	2			
2	Prol	blemfälle	3			
	2.1	Technische Semantik	3			
		2.1.1 Sonderzeichen	3			
		2.1.2 Leerzeichen	4			
		2.1.3 Zeilenbrüche	6			
		2.1.4 Dokumentenbrüche	7			
	2.2	Sprachliche Semantik	8			
		2.2.1 Ausgangssituation	8			
		2.2.2 Exemplarische Beispiele	8			
3	Tecl	hnologien	9			
	3.1	Übersicht	9			
		3.1.1 Auflistung	9			
		3.1.2 Eingrenzung	9			
		3.1.3 Auswertung	9			
	3.2	Einschätzung	9			
	3.3	Fazit	9			
,	Offene Problematiken					
4			10			
	4.1	Verfolgte Ideen	10			
	4.2	Gelöste Probleme	10			
	4.3	Lessons Learned	10			
	4.4	Fazit	10			
5	Fazi	it	11			
	5.1	Zusammenfassung	11			
	5.2	Ausblick	11			
	5.3	Weiteres	11			
6	Eige	enständigkeitserklärung	12			
Lit	eratı	ur	13			
Α		änge	14			
	A.1	Fontskalierung auf Webseiten	14			

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)"). Software zur Übersetzung von menschlicher Sprache auf TEX-Quellcode anzuwenden, kann Dokumente erzeugen, welche die TFX-Syntax brechen oder nur zu Teilen und damit unvollständig übersetzt wurden. Mit Hilfe von Google Translate lassen sich wesentliche Gründe ermitteln und welche weitere Folgen aus einem Übersetzungsprozess entstehen können, der nicht von einem LATEX-Dokument ausgeht (mit: TEX, der Programmiersprache für typesetting (Zeichensetzung) von Knuth (1986) sowie der Erweiterung des funktionellen Umfangs durch Lamport (1994)).Bereits einfachste Funktionen dieses Systems können Übersetzungen einfacher Zeichenketten verhindern. Ein Ubersetzen (via Google Translate) von hello wor\textit{ld} liefert nicht Hallo We\textit{lt}, sondern hallo wor\textit{ld}. Abgesehen von der Frage, wo die kursive Hervorhebung im eigentlichen String erfolgen sollte, würden Leser eines kompilierten Dokumentes das Wort "Welt" erkennen.Die beschriebene Zeichenkette wird von TEX als "hello world" dargestellt, in welcher das Wort "world" für einen menschlichen Leser als das englische Wort für "Welt" eindeutig erkennbar ist. Fehlt die Kenntnis über eine der Sprachen (DE,EN) oder ein natürliches Sprachverständnis verliert die Wortkette einen Teil ihrer Bedeutung. Besonders fatal wird 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 met 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.

¹Zwei 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 Technische Semantik

2.1.1 Sonderzeichen

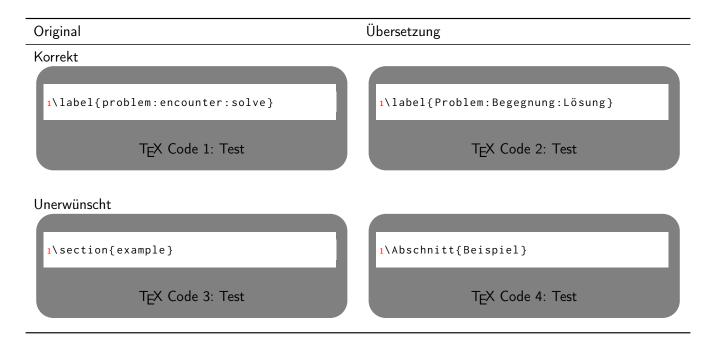


Tabelle 1: Fehler in einem Token

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 erkennt².

Takeaway Teile der TEX-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

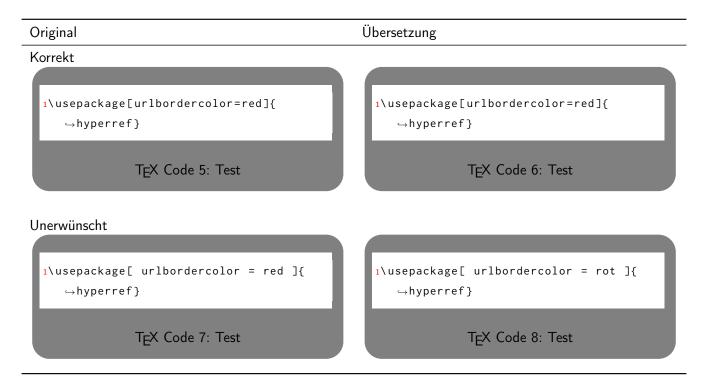


Tabelle 2: Fehler in einem einzeiligen Dokument

²,Vernestung" meint die Verschachtelung von Klammern

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

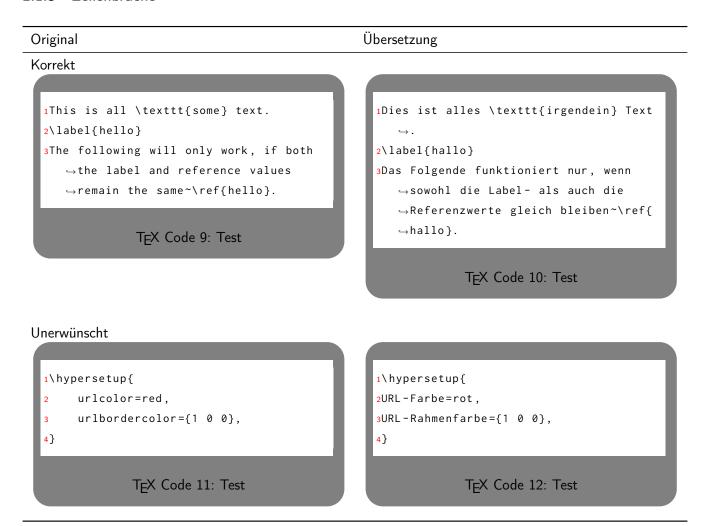


Tabelle 3: Fehler in einem einzeiligen Dokument

Beschreibung

Abstrahierung Teile der T_EX-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

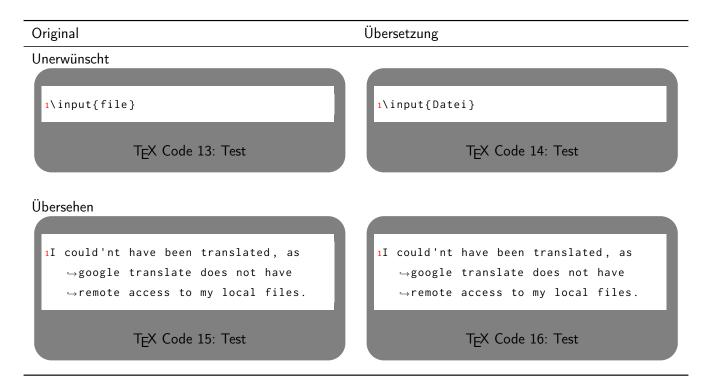


Tabelle 4: Übersetzung in einem include führt zum nicht-Übersetzen einer Datei

Beispiel

Beschreibung

Abstrahierung Teile der TEX-Syntax müssen nicht zwingend in einer Datei vorliegen, sondern könnten auch in verschiedenen Dateien integriert sein. Die Klassifizierung technischer 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 formulierbar³ und äußert sich in diesem Kontext auf eine Erhöhung von Laufzeitkomplexitäten.

³physikalisch: die vierte Dimension ist die Zeit, wenn man eine nicht-euklidische 3-dimensionale Bewegung verlangt (Teleportation)

2.2 Sprachliche Semantik

2.2.1 Ausgangssituation

Probleme des vorherigen Teils 2.1 zu verhindern, würde ein Kompilieren eines übersetzten Dokumentes gewähren, als auch keine Inhalte von diesem übersehen. Diese sollten als minimale Anforderung zu sehen sein. Die Fähigkeit alle Inhalte eines LATEX Dokumentes lesen zu können, zeigt sich unausreichend, da das Übersetzen zwischen menschlichen Sprachen⁴ je nach Kontext andere Lexeme produzieren muss⁵. Dieser Kontext stammt in einem Quellcode aus mehr als nur den rein wörtlichen Inhalten und kann auch aus bestimmten Elementen eines Dokumentes hervorgehen. Ein Bestimmen des *genauen* Kontextes erfordert Kenntnis über diese *bestimmten Elemente* eines Dokumentes, welche alle ihrerseits verschiedenst charakterisiert sein könnten und werden könnten. Es ist nicht garantierbar, dass die Menschheit für die gesamte Zukunft den selben Paradigma folgen wird, mit welchen sie heute Informationen veranschaulicht. Als Beispiel eine Frage: Sahen alle Bilder von Menschen des Steinzeitalters gleich aus?

Vorangegangenes zeigt demnach eine unendliche Menge erwartbarer Probleme. Um innerhalb der Grenzen der Realität zu verweilen, wird sich auf Anwendungsfälle von LATEX für z.B. wissenschaftliche Arbeiten oder Veröffentlichungen konzentriert und inwiefern durch die hierfür benötigten Methoden (bspw. Zitationen, Tabellen, Formeln, Graphiken,...) Kontexte nicht aus der sprachlichen Satzstellung alleine hervorgehen.

2.2.2 Exemplarische Beispiele

Autoren/Titel Ein Autor wird vermutlich innerhalb eines Dokumentes seinen eigenen Namen oder den Titel des jeweiligen Werkes nicht öfters erwähnen, aber aus diesem lassen sich unter Umständen mehr Informationen über den Kontext gewinnen. Handelt es sich um einen bekannten Autoren könnte der Name einen Kontext liefern (welcher jedoch im Vorwege nur als eine Heuristik angesehen werden kann, da der Autor authentifiziert werden muss) oder der Titel trägt den Namen von einer bekannten Methode aus einem wissenschaftlichen Gebiet oder nennt ein solches Gebiet. Denkbare Beispiele umfassen: "[...] of Communication", "Bayesian [...]", "[...] inference [...]" (entnommen aus einigen subject areas von ACM: Transactions on Probabilistic Machine Learning). Vergisst aber ein übersetzendes Programm diese Wörter des Titels an einer kritischen Stelle im Text, so würden fälschliche Übersetzungen entstehen.

Dokumenten-interne Verweise

⁴Formulierung "menschliche Sprachen": umfasst alle menschlichen Kommunikationswege, auch sprachenunabhängige non-Verbale (bspw. mathematische Notationen) und bildliche Wörter (bspw. chinesische Hanzi, japanische Kanji, theoretisch gesehen auch Gebärdensprachen)

⁵Das Übersetzen aus und in Gebärdensprachen ist vorerst nicht im Fokus. Denkbarer Ansatz wäre hier allerdings Dokumente mit Bildern zu erwarten, bei welchen man davon ausgeht, dass diese mit bekannten Methoden der Mustererkennung erkannt und in Wörter (wie sie in den lateinischen Sprachen verwendet werden) überführt werden können und dann maschinell ausgewertet werden können. Hierzu gibt es zahlreiche Technologien (Rastgoo et al. 2025), jedoch beschäftigt dies ein Übersetzen aus einer menschlichen Sprache nach LATEX und weicht damit von der gegebenen Aufgabenstellung ab. (Obwohl auch hierfür Ansätze existieren: Korzh et al. (2025))

Externe Referenzen Zwar ist das Nutzen einer Technologie wie BibTeX zur

Formeln, Tabellen

Graphiken

2.3 Spezifischer Technologien

Hier wenden wir uns von Problemen einer Übersetzung ab und widmen uns denen eines Lesers. Alle textlichen Inhalte eines Dokumentes zu übersetzen, als auch eine kontextuelle Fachsprache zu bewahren scheint aus abstrakterer Perspektive ausreichen, kann allerdings zu Situationen führen, in welchen Informationen verloren gehen, da diese vom Endnutzer nicht mehr gesehen werden können.

2.3.1 Kommentare

Beispiele

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

Abstrahierung Hier treffen technische 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 ?? beinhalten.

2.3.2 Dilemmatische Makros

Beispiele

Beschreibungen

Abstrahierung

2.3.3 TikZ und Layouting

Beispiele

Beschreibungen

Abstrahierung

~ ~ 4	\sim 1					٠.
2.3.4	Quel	Imar	rcni	'ack	אות	VOIT
2.3.4	Quei		II SPI	acı	II g	veir

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 Weitere Schwierigkeiten

Kommentare

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	Eigenst	ändig	keitserk	lärung
U	Ligerist	anuig	NCILSCI N	iai ang

_	_	_		
habe. Dazu habe i benutzten Werken	ich keine außer den inhaltlich und wört	von mir angegebenen :lich entnommenen St	tändig angefertigt und oh Hilfsmitteln und Quellen ellen habe ich als solche en gedruckten Exemplaren	verwendet und die den kenntlich gemacht. Ich

Rostock, den 02.12.2025

Hendrik Theede

Literatur

- Knuth, D. E. (1986), The TeXbook, ISBN: 9780201134476, Addison-Wesley Professional.
- Korzh, D., Tarasov, D., Iudin, A., Karimov, E., Skripkin, M., Kuzmin, N., Kuznetsov, A., Rogov, O. Y. & Oseledets, I. (2025), 'Speech-to-latex: New models and datasets for converting spoken equations and sentences', https://doi.org/10.48550/arXiv.2508.03542 (last Access: 06.10.2025).
- Lamport, L. (1994), LaTeX: A Document Preparation System, 2nd Edition, ISBN: 9780201529838, Addison-Wesley Professional. available at: https://www.latex-project.org/help/books/tlc3-digital-chapter-samples.pdf (last Access: 04.10.2025).
- Rastgoo, R., Kiani, K. & Escalera, S. (2025), 'Sign language recognition: A deep survey', https://doi.org/10.1016/j.eswa.2020.113794 (last Access: 06.10.2025).

A Anhänge

A.1 Fontskalierung auf Webseiten

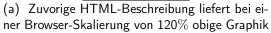
Beispielsweise produziert die folgende HTML-Notation bei einer Skalierung im Browser von 120 Prozent (Abbildung 2a) und 50 Prozent (Abbildung 2b) jeweilig zwei verschiedene PDF (unter welchen nur Zweitere alle textlichen Inhalte offenbart). Ähnliches kann auch innerhalb T_FX geschehen, sollte

```
<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