

Entwurf und Implementierung einer Werkzeugunterstützung zur sprachlichen Analyse und automatisierten Transformation von Projektlastenheften im Kontext der Automobilindustrie

An der Fachhochschule Dortmund

im Fachbereich Informatik

Studiengang Informatik

Vertiefung Praktische Informatik

erstellte Thesis

zur Erlangung des akademischen Grades

Bachelor of Science

B. Sc.

von Aaron Schul,

geboren am 24.06.1997

und Felix Ritter

geboren am 31.08.1997

Betreuung durch:

Prof. Dr. Sebastian Bab und Prof. Dr. Steffen Helke

Dortmund, 28.02.2019

Zusammenfassung

Ein weitläufiges Problem in der Autoindustrie ist die effiziente Verarbeitung von Pflichtenheften. Einer der Gründe dafür ist, dass viele verschiedene Bereiche bei der Entwicklung der Fahrzeuge und sogar kleinster Einzelteile beteiligt sind. So müssen zu Beginn einer Produktentwicklung etwa Betriebswirtschaftler, Designer, Techniker und Ingenieure zusammen ein Dokument entwerfen, das die Produktmerkmale widerspiegelt. Darin müssen die Anforderungen an das gewünschte Produkt so genau beschrieben sein, sodass es anhand dieses Dokuments entwickelt werden kann. Als Experten ihrer jeweiligen Domäne weiß dabei jeder genau, was dazu nötig ist. Sobald jedoch Auswirkungen über die eigene Domäne hinausgehen, kann es schnell passieren, dass Widersprüche oder Abhängigkeiten entstehen. Diese werden später schnell übersehen, da solche Zuständigkeiten nicht eindeutig geklärt sind. Das Resultat ist dann ein Fehler in der Entwicklung, welcher von zeitlicher Verzögerung über zusätzliche Kosten bis hin zum Abbruch des Projekts führen kann. Methoden aus dem Natural Language Processing (NLP) können dabei helfen, die Überprüfung der Pflichtenhefte auf Konsistenz stark zu vereinfachen oder sogar teilweise zu automatisieren. Dafür kann beispielsweise der Text in den Lastenheften analysiert und dessen Inhalt innerhalb einer Wissensbasis, sog. Ontologie, abgespeichert werden. In dieser Arbeit wird daher die Entwicklung zweier NLP-basierter Werkzeuge zur Vereinfachung bzw. Lösung dieser Probleme vorgestellt werden. Zum einen ein *Requirements-to-Boilerplate-Converter*-Werkzeug (R2BC), welches dem Nutzer helfen soll, das Pflichtenheft eines Auftraggebers in die betriebsinternen Richtlinien und Standards zu überführen. Zum anderen wird der *Delta-Analyser* vorgestellt, welcher auf dem R2BC aufbaut, indem er automatisch zwei homogene Lastenhefte vergleicht und dadurch im Kontext des gesamten Pflichtenhefts Widersprüche und Abhängigkeiten herausstellt. Ziel ist es dabei, jeweils einen Prototypen als „*proof of concept*“, also als Machbarkeitsstudie, zu implementieren. Nach dieser Vorstellung wird zusätzlich das weitere Potential der Programme erläutert, welches sich bei der Entwicklung der Prototypen gezeigt hat. Diese bietet Vorschläge zur weiteren bzw. vollständigen Implementierung der Programme.

Abstract

A widespread problem in the auto industry is the efficient processing of specifications. One of the reasons for this is that many different areas are involved in the development of vehicles and even the smallest of parts. For example, at the beginning of product development, managers, designers, technicians, and engineers all need to design a single document together. The requirements for the desired product must be described in such detail that it can be developed using this document. As experts in their domain, everyone knows exactly what is needed. However, as soon as effects go beyond the own domain, it can quickly happen that contradictions or dependencies arise. These are quickly overlooked later, as such responsibilities are not clearly clarified. The result is then a mistake in the development, which can lead from time delay over additional costs up to the demolition of the project. Methods from Natural Language Processing (NLP) can help to simplify or even partially automate the review of functional specifications for consistency. For example, the text in the specifications can be analyzed and its contents stored within a knowledge base, so-called ontology. In this work, therefore, the development of two NLP-based tools to simplify or solve these problems will be presented. On the one hand, there is a program called Requirements-to-Boilerplate Converter (R2BC), which should help the user to transfer the specifications of a client into the company's internal guidelines and standards. On the other hand, the so-called delta analyzer is shown, which is based on the R2BC, in that it automatically compares two homogeneous specification sheets and thereby highlights contradictions and dependencies in the context of the entire specification. The goal here is to implement a prototype as a proof of concept of the program concepts from ZIC19. Due to the limited resources of this work, a list with further potential of the programs, which has been shown during the development of the prototypes, is additionally explained. This offers suggestions for further or complete implementation of the programs.

Inhaltsverzeichnis

1	Einführung	9
1.1	Motivation	9
1.2	Unternehmenskontext	12
1.2.1	Einsatz bei HELLA	12
1.2.2	Anforderungsmanagement im Unternehmen	12
1.3	Hypothese	13
1.4	Methodik	14
1.5	Aufbau der Arbeit	14
1.6	Autorenverzeichnis	15
2	Stand der Technik	16
2.1	Rückblick auf die PA	16
2.1.1	Grundlagen zu Natural Language Processing	16
2.1.2	Umsetzung der Verarbeitungsschritte natürlicher Sprache	19
2.1.3	Syntaktische Analyse	21
2.1.4	Umgang mit Semantik	26
2.2	Ontologien	27
2.2.1	Definition	28
2.2.2	Aufbau	28
2.3	Verwandte Arbeiten	32
3	Betriebliches Umfeld - Use-Case HELLA	36
3.1	Allgemeines zu HELLA	38
3.2	RE bei HELLA	39
3.2.1	Abteilung für RE	39
3.2.2	RE-Prozess	40
3.3	Anforderungen im RE-Management	43

3.3.1	Kriterien der Qualität von Anforderungen	43
3.3.2	Kriterien zur Herstellung qualitativer Anforderungen .	47
3.3.3	Umgang mit natürlicher Sprache in Lastenheften . . .	48
3.4	Anwendungsfall bei HELLA	48
3.5	Ansatz und Konzept	49
3.5.1	Anforderungen an die Entwicklung	50
3.5.2	Zuordnung einzelner Werkzeuge	50
3.5.3	R2BC-Funktionalität	53
3.5.4	DA-Funktionalität	56
4	Requirements-To-Boilerplate-Converter (R2BC)	60
4.1	Architektur, Klassen und Verteilung der Ressourcen	60
4.1.1	Verwendete Software	61
4.2	Klassenstruktur	62
4.3	Werkzeug-Implementierung	66
4.3.1	Sprachverarbeitung mit GATE	66
4.3.2	Converter-Funktionalität	79
4.3.3	GUI - Nutzerinteraktion und Workflow	83
4.4	Mögliche Erweiterungen	83
4.4.1	Benutzeroberfläche	84
4.4.2	Nutzbarkeit	85
4.5	Test	88
4.5.1	Methodik	88
4.5.2	Durchführung	88
4.5.3	Ergebnisse	88
5	Delta-Analyse	89
5.1	Architektur	89
5.2	Implementierung	89
5.3	mögliche Erweiterungen	89
5.4	Test	89
5.4.1	Methodik	89
5.4.2	Durchführung	89
5.4.3	Ergebnisse	89
6	Evaluation	90
6.0.1	Umsetzung der NLP-Architektur in unserem Werkzeug	90
6.1	Auswertung der Testresultate	93

6.2	Mehrwert?	93
6.3	Ziel erreicht? Hypothese reviewen und schwafeln	93
7	Fazit	94
7.1	Zusammenfassung	94
7.2	Ausblick	94
8	Notizen	95

Abbildungsverzeichnis

2.1	Nicht-deterministischer Zustandswandler der able-Regel [RS18]	23
3.1	Überblick der 3 Qualitätsstufen von Anforderungen im <i>Requirements-Management & Engineering</i> (RM & E) gemäß [HP12]	47
3.2	Szenario für Unterstützung durch R2BC-Komponente [eigene Darstellung]	52
3.3	Szenario für Unterstützung durch DA-Komponente [eigene Darstellung]	53
3.4	Verteilung der R2BC-Ressourcen [eigene Darstellung]	55
3.5	Verteilung der DA-Ressourcen [eigene Darstellung]	58
4.1	Zuordnung der R2BC-Klassen zu Paketen nach Abb.3.x [eigene Darstellung]	63
4.2	Klassendiagramm R2BC [eigene Darstellung]	63
4.3	Anforderung mit farbig markierten Token-Annotationen in GATE	67
4.4	Ausgewählte Token-Annotation und dessen Features in GATE	68
4.5	Verwendete ANNIE-Verarbeitungsressourcen in GATE [eigene Darstellung]	69
4.6	Verarbeitungslogik aus ANNIE und JAPE Transducern [TOL09]	70
4.7	Beispielhafte JAPE-Regel [eigene Darstellung]	71
4.8	Ausschnitt der <i>SystemName</i> -Liste des Annie-Gazetteers [eigene Darstellung]	74
4.9	Definitionsdatei des Annie-Gazetteers [eigene Darstellung]	75
4.10	JAPE-Regel zur Erkennung von eingeschränkten Akteuren [eigene Darstellung]	75
4.11	JAPE-Regel zur Markierung von Sätzen für Boilerplate ⁸⁵ [eigene Darstellung]	77

4.12	Alle Features eines <i>Boilerplate85</i> -Satzes in GATE [eigene Darstellung]	78
4.13	Initialisierung der GATE-Ressourcen in <i>initFromRestore()</i> -Methode der Klasse <i>AnnieHandler</i> [eigene Darstellung]	81
4.14	Ausführung der ANNIE-Pipeline, Verarbeitung der Annotationen in <i>BoilerplateHandler</i> in <i>convert()</i> -Methode der Klasse <i>AnnieHandler</i> [eigene Darstellung]	82
4.15	JAPE und JAPE+ Zeitaufwand abhängig von der Dokumentlänge	87

Tabellenverzeichnis

Abkürzungsverzeichnis

CRS	Customer Requirement Specification
DA	Delta-Analyser
E-AE	Electronics-Advanced Engineering
NLP	Natural Language-Processing
OEM	Original Equipment Manufacturer
OWL	Web-Ontology-Language
PMT	Projects, Methods and Tools
POS	Part-Of-Speech
R2BC	Requirements-To-Boilerplate-Converter
RE	Requirements-Engineering

Danksagung

An dieser Stelle möchten wir uns bei allen Personen bedanken, die uns bei unserer Bachelorarbeit und während des Studiums begleitet und unterstützt haben.

Der Dank gebührt unseren Familien und Freunden aus der Heimat und aus dem Studium für ihre Unterstützung.

Wir danken auch allen Kollegen aus der Abteilung E-AE bei der Firma HELLA in Lippstadt für ihr Interesse und die großartige Zusammenarbeit. Die einzigartigen Erfahrungen, die wir während unseres Praktikums und später während der Bachelorarbeit in der Firma sammeln konnten, haben uns wirklich motiviert. Besonderer Dank geht dabei an Konstantin Zichler, der weit mehr als nur Betreuer in der Firma, sondern auch Mentor und Ideengeber unserer Arbeit war.

Schlussendlich danken wir unseren Betreuern Prof. Dr. Sebastian Bab und Prof. Dr. Steffen Helke für die zahlreichen Tipps und Anregungen während der Bachelorarbeit.

Kapitel 1

Einführung

1.1 Motivation und thematische Grundlagen

In einer Vielzahl von Firmen stellt die Erhebung von Anforderungen den ersten Schritt bei der Entwicklung von neuen Produkten dar. Diese Anforderungen manifestieren sich später in der zu entwickelnden Hard- und Software. Spätere Produktmerkmale müssen dabei weit vor der tatsächlichen Entwicklung berücksichtigt werden und fließen in das Anforderungsdokument ein. Die Erhebung und Identifikation dieser Anforderungen wird als *Requirements-Engineering* bezeichnet [BAL10].

Dieser Prozess wird dabei sowohl bei kompletten Neuentwicklungen, als auch bei der Adaption eines bereits vorhandenen Produktes auf neue Projekte durchlaufen. Zur Rolle von Requirements-Engineering in Projekten siehe etwa [MW02] und [HL01].

Dem Betrieb liegen anschließend die Aufgaben und Ziele für die Entwicklung als Dokumente vor, auf deren Basis die Produktentwicklung beginnen kann. Die Entwicklung unterliegt dabei bestimmten Kriterien und Faktoren, die den unternehmerischen Erfolg beeinflussen. Neben betriebswirtschaftlichen Einflüssen wie der Einordnung des Produktes in der Wertschöpfungskette sind es dabei besonders technische Anforderungen an das Produkt, die definiert und während der Produktentwicklung eingehalten werden müssen.

Anforderungen im betrieblichen Entwicklungsprozess

Das Resultat ist das sogenannte Pflichtenheft, in Unternehmen häufig als CRS (*Customer Requirement Specification*) bezeichnet [DGE19]. Mithilfe des-

sen kann nun die Entwicklung eines Produktes begonnen werden. Ein später hergestelltes Produkt sollte möglichst allen beschriebenen Anforderungen entsprechen. Da Betriebe jedoch nur selten alle einzelnen Komponenten des späteren Produktes selbst herstellen können, sind sie auf Zulieferer angewiesen, die einzelne Komponenten entwickeln und produzieren. Diese Hersteller einzelner Komponenten werden als *OEM* (*Original Equipment Manufacturer*), zu Deutsch *Erstausrüster*, bezeichnet [IR09].

In diesem Fall wird dem Zulieferer ein Pflichtenheft übergeben, dass allen Anforderungen des Kunden genügt. Das bedeutet jedoch, dass auch die Produktentwicklung anhand des Pflichtenheftes an den Zulieferer ausgelagert wird. Dem Zulieferer werden damit Möglichkeiten geboten und Verantwortung übertragen, das Produkt weiter zu definieren. Einerseits können etwa interne Richtlinien und Erfahrungen auf dem Gebiet der Produktentwicklung aus früheren Projekten integriert werden, andererseits aber auch weitere Optimierungen sowie Merkmale, die noch nicht im Pflichtenheft dokumentiert sind. Somit agiert der Zulieferer nicht nur als produzierende Fabrik, sondern nimmt am Entwicklungsprozess teil.

Die Anforderungen an ein Produkt, etwa technische Rahmenbedingungen und Qualitätsanforderungen, setzen sich dabei aus einer Auflistung von funktionalen und nicht-funktionalen Anforderungen zusammen [BAL10]. Häufig enthält das Pflichtenheft auch weitere Informationen und abseits von Text Abbildungen, die einzelne Anforderungen ergänzen.

Umstände der Anforderungsdefinition

Die Anforderungen werden dabei von vielen verschiedenen Domänenexperten beim OEM formuliert und in das Pflichtenheft für die Entwicklung eingetragen. Verschiedenste Akteure aus einem Unternehmen sind dabei an der Festlegung der Anforderungen an ein Entwicklungsprojekt bzw. Produkt beteiligt. Beteiligte sind etwa Produktdesigner, Ingenieure und Systemtechniker, die an verschiedenen Stellen im Lastenheft Anforderungen an eine Komponente festlegen und an entsprechenden Stellen vermerken. Diese Beteiligten sind in der Regel auf ihren Bereich spezialisiert und nicht interdisziplinär, da häufig an verschiedenen Orten und an spezifischen Aspekten eines Produktes in mehreren Projektteams entwickelt wird.

Projektteams aus einem Bereich tauschen sich häufig nicht untereinander aus und legen unabhängig von anderen Beteiligten Anforderungen fest. Demzufolge sammeln sich im Lastenheft schon bei der Entstehung beim OEM

verschiedenste Merkmale einer Komponente, die aber nicht im Bezug zueinander stehen und sich im schlimmsten Fall gegenseitig ausschließen. Ferner gibt es sprachliche Eigenheiten der Autoren und sprachliche Fehler, die innerhalb des Teams nicht auffallen und später nicht mehr korrigiert werden. Auch gibt es unternehmensinterne Richtlinien für die Formulierung von Anforderungen beim OEM, die das Verständnis auf Seite des Zulieferers erschweren können. Christian Allmann, Lydia Winkler und Thorsten Kölzow haben in [AWK06] solche Herausforderungen für das Requirements-Engineering zwischen OEMs und Zulieferern identifiziert.

Durch die entstehende große fachliche Breite und Tiefe der Spezifikationen im Pflichtenheft, können mitunter Dokumente mit einem Umfang von von mehreren tausend Seiten entstehen. Dies bedeutet einen hohen Aufwand bei der Transformation und Bündelung der Anforderungen in ein tatsächliches Produkt beim Zulieferer.

Prozessoptimierung durch Werkzeuge

Bei der Untersuchung von Projektlastenheften muss also nach Zusammenhängen und Bezügen zwischen mehreren Anforderungen gesucht werden, damit die Korrektheit des späteren Produktes gewährleistet ist. Die Analyse von Zusammenhängen zwischen Anforderungen stellt dabei aus Gründen der Effizienz ein Problem dar, wenn jeder Projektbeteiligte von Hand die für ihn relevanten Anforderungen aus dem Lastenheft extrahieren muss. Auch müssen die Lastenhefte an die Formulierungen und Ausdrucksweisen für Requirements-Management im Unternehmen angepasst werden.

Bislang gibt es jedoch kaum Werkzeugunterstützung, die effiziente Möglichkeiten zur automatisierten Überarbeitung und Anpassung einzelner Anforderungen aus dem Dokument bietet. Ansätze aus dem *Natural-Language-Processing*, kurz NLP, stellen gleichzeitig vielversprechende Forschungsfelder in der Informatik dar, die eine solche automatisierte Verarbeitung auf Basis von Sprachanalyse ermöglichen. Diese werden in Kap. 2 ausführlich erläutert. Syntax und Semantik der einzelnen Sätze und Zusammenhänge in Texten können auf Basis aktueller Trends wie Machine-Learning und dynamischer Programmierung zunehmend besser abgebildet werden.

1.2 Unternehmenskontext der Arbeit

Besonders betroffen von dem geteilten Entwicklungsprozess zwischen OEMs und Zulieferern ist die Automobilindustrie. Eine Vielzahl von Zulieferern beliefert die Automobilhersteller mit Komponenten für ihre Fahrzeuge. Einzelne Komponenten bestehen dabei häufig auch aus mehreren Einzelteilen, die wiederum von mehreren verschiedenen Zulieferern geliefert werden. [AWK06]

1.2.1 Einsatz bei HELLA

Einer dieser Zulieferer ist die HELLA GmbH & Co. KGaA (im folgenden (die) HELLA genannt), die sich auf die Entwicklung und Produktion von Licht- und Elektronikkomponenten in Hard- und Software spezialisiert hat. HELLA ist ein international operierender deutscher Automobilzulieferer mit Hauptsitz in Lippstadt und zählt zu den Top 100 Automobilzulieferern weltweit [LV10]. Für weitere Information siehe [HE19] sowie die Firmenwebsite.

Zum Verfassen dieser Arbeit waren die Autoren fünf Monate bei der Firma beschäftigt. Dies beinhaltete zwei Monate Arbeit als Praktikanten und die verbleibenden drei Monate als Bacheloranden. Die Hauptaufgabe dort war die Mitarbeit im Bereich Requirements-Engineering für die Abteilung der Elektronik-Vorentwicklung *E-AE* (*Electronics-Advances Engineering*).

Die Abteilung ist zuständig für die Planung und Umsetzung von Vorentwicklungsprojekten um die Ergebnisse an die Serienentwicklung oder verschiedene OEMs weiterzuleiten. Dies beinhaltet häufig, wie zuvor beschrieben, den Austausch von Informationen der Firma HELLA mit diversen OEMs bezüglich gemeinsamer Projekte. Die genaue Darstellung dieser Prozesse erfolgt in Kap. 3.1. Somit erhält diese Abteilung auch oft Lastenhefte der OEMs, welche dann an die Projektteams weitergeleitet werden.

1.2.2 Anforderungsmanagement im Unternehmen

Die Gewichtung einzelner Anforderungen in einem größeren Systemkontext fällt dort schwer, da nun Projektteams beim Zulieferer, die an der Entstehung des Lastenheftes nicht beteiligt waren, dieses verstehen sollen und auf ihre Teilkomponente anwenden müssen. Zusätzlich zu den teilweise widersprüchlichen Anforderungen aus dem Pflichtenheft kommen weitere Anforderungen des Zulieferers, wie etwa die Gewinnmarge und unternehmensinterne Richtlinien hinzu. Schlussendlich soll jedoch ein Produkt entwickelt werden, dass

möglichst alle Anforderungen berücksichtigt. [MW02]

Zu diesem Zweck können bei Zulieferern spezielle Abteilungen existieren, die Projektteams bei der Identifikation und Gewichtung von Anforderungen unterstützen. Diese Abteilungen sind auf die Auswertung der Projektlastenhefte spezialisiert und erstellen Modelle zu Merkmalen des späteren Produktes. Die Experten stellen dabei die Eindeutigkeit und Verständlichkeit sicher und identifizieren relevante Anforderungen für die Mitglieder des Projektteams. Somit kann ein eigenes Anforderungsdokument für das Projektteam beim Zulieferer erstellt werden. Falls nicht vorhanden, muss das Team selbst über die Bewertung von Anforderungen entscheiden und Rücksprache mit dem OEM halten.

Bei HELLA existiert eine dementsprechende Abteilung, genannt *Projects, Methods and Tools*, kurz *PMT*. Im Rahmen unserer Beschäftigung konnte in enger Zusammenarbeit mit PMT genau ermittelt werden, wie Lastenhefte momentan verarbeitet werden, welche Verbesserungspotentiale dabei bestehen und wie diese genutzt werden können.

1.3 Hypothese

Hypothese dieser Arbeit ist, dass sich mithilfe von NLP Lastenhefte effizient automatisiert verarbeiten lassen. Damit kann die Arbeit von Requirements-Engineers, aber auch von Beteiligten an der Entwicklung beim Verständnis der Anforderungen, erleichtert werden. Dabei ist die Auswertung der Syntax und Semantik für ein tieferes Verständnis von Textzusammenhängen, also von verschiedenen Anforderungen, relevant.

Die sprachliche Analyse kann Requirements erkennen und kategorisieren, sowie in Beziehung zueinander setzen. Insbesondere lassen sich die beschriebenen Probleme bei der Überprüfung der Pflichtenhefte auf Konsistenz lösen oder zumindest vereinfachen. Dies kann mithilfe der in dieser Arbeit beschriebenen Programme geschehen, welche konzeptuell dargestellt und prototypisch implementiert werden.

Weiterhin soll geprüft werden, wie stark ein Mehrwert bei der Nutzung solcher Programme für ein Firma ausfallen kann. Dies beinhaltet ersparte Arbeitszeit, Kosten und vermiedene Fehler.

1.4 Methodik

Verschiedene Methoden aus dem Bereich Natural Language-Processing werden zunächst anhand des Rückblicks auf die Projektarbeit theoretisch und anhand praktischer Beispiele erläutert. Der betriebliche Kontext der Firma HELLA wird durch genaue Darstellung der dortigen Verfahren beleuchtet.

Mit diesen Grundlagen wird dann die Unterstützung der dortigen Prozesse im Bereich Requirements-Engineering auf Basis von natürlichsprachlicher Textanalyse evaluiert. Die gewonnenen Erkenntnisse aus der Zusammenarbeit mit der Abteilung PMT wurden in Anforderungen an die zu entwickelnden Werkzeuge zusammengefasst, die als Basis für unsere Softwareentwicklung dienten. Dadurch entstand eine genaue Vorstellung der zu entwickelnden Software, welche mit Hilfe der Programmkonzepte von Konstantin Zichler ZIC19 entworfen und implementiert werden konnte.

Das Ergebnis stellten zwei Software-Prototypen auf *java*-Basis dar, deren Struktur und Funktionsweisen wir anhand von Modellen und Codebeispielen darstellen. Die Prototypen wurden mit originalen Pflichtenheften von verschiedenen OEMs und der HELLA getestet und evaluiert.

1.5 Aufbau der Arbeit

In der folgenden Ausarbeitung wird zunächst in Kapitel 2 ein Überblick über den Stand der Technik gegeben. Dies beinhaltet einerseits einen Rückblick auf die zuvor verfasste Projektarbeit zu NLP und einer Erläuterung der Grundlagen von Ontologien und andererseits einen Überblick von Arbeiten, die sich der gleichen oder ähnlichen Problemstellungen gewidmet haben.

Kapitel 3 beschäftigt sich für eine bessere Einordnung der beiden Programme mit dem betrieblichen Umfeld bei der HELLA. Hierzu wird zunächst, vor dem Hintergrund des momentanen Vorgehens, der Anwendungsfall der Programme genauer beschrieben. Somit lassen sich die betrieblichen Anforderungen an die Werkzeuge bestimmen, welche dann von den vorgestellten Konzepten abgedeckt werden sollen.

In Kapitel 4 wird das erste Werkzeug - der R2B-Converter - ausführlich beschrieben. Dies beinhaltet die Darstellung dessen Architektur, eine Vorstellung der Implementierung, eine Auflistung möglicher Erweiterungen und eine genaue Beschreibung des Testverfahrens der Software. Zu den Tests werden dabei präzise die Methodik, die Durchführung und die Ergebnisse

vorgestellt, welche in dem späteren Evaluations-Kapitel aufgegriffen und eingeordnet werden.

Das zweite Werkzeug - der Delta-Analyser - wird in Kapitel 5 behandelt. Dabei werden ähnlich zu Kapitel 4 die Architektur, die Implementierung, mögliche Erweiterungen und durchgeführte Tests beschrieben. Es wird genau dargelegt wie die Testergebnisse erzielt wurden, die später zusammen mit den Ergebnissen von Kapitel 4 evaluiert werden, um das Resultat der beiden Werkzeuge bewerten zu können.

Kapitel 6 stellt das zuvor erwähnte Evaluations-Kapitel dar. Hier findet zunächst die Auswertung der erhobenen Testergebnisse statt. Anschließend wird anhand dessen dargestellt, ob und inwiefern das Nutzen der Werkzeuge einen Mehrwert für die Firma erwirtschaften kann und somit im betrieblichen Umfeld eingesetzt werden sollte. Somit lässt sich die Korrektheit der in 1.2 gestellten Hypothese überprüfen.

Die Arbeit wird von Kapitel 7 abgeschlossen. Dieses beinhaltet eine Zusammenfassung der geleisteten Arbeit und der Ergebnisse. Außerdem wird zuletzt noch ein Ausblick für die Zukunft der Werkzeuge und die Lösung des Problems erläutert.

1.6 Autorenverzeichnis

1.: Ritter, Schul

1.1: Ritter

1.2: Ritter

1.3: Ritter, Schul

1.4: Schul

1.5: Schul

2.: Ritter, Schul

2.1: Schul

Kapitel 2

Stand der Technik

In diesem Kapitel wird zunächst ein Rückblick auf die zuvor von den Autoren verfasste Projektarbeit und die darin enthaltenen Grundlagen von NLP gegeben. Aufgrund ihrer besonderen Relevanz für die Programme werden anschließend zusätzliche Ontologien als Teil von NLP genauer beleuchtet. Außerdem wird mit Hilfe von verwandten Arbeiten ein Überblick über die Problematik und verschiedene Lösungsansätze geboten.

2.1 Rückblick auf die Projektarbeit

In [RS18] sind die Grundlagen von NLP im Rahmen einer Projektarbeit zusammengefasst. In diesem Abschnitt werden die für die Arbeit relevanten Aspekte kurz wiederholt.

2.1.1 Grundlagen zu Natural Language Processing

Definition und Ziel

Natural Language Processing ist zu verstehen als die automatisierte oder halb-automatisierte Verarbeitung von natürlicher Sprache mit Hilfe eines Computers. In NLP sind verschiedenste wissenschaftliche Bereiche verknüpft.¹ Hauptsächlich handelt es sich dabei um Linguistik und Informatik, jedoch

¹Es ist anzumerken, dass verschieden weit gefasste Definitionen von NLP existieren. Manche schließen etwa auch die Erfassung von Sprache mit ein, etwa durch Interaktion mit Sprachcomputern. Andere beschränken sich ausschließlich auf den Verarbeitungsprozess der Sprachdaten im Computer selbst, siehe 2.1.2.

gibt es auch viele Verbindungen zur Psychologie, Philosophie Mathematik und Logik. [COP04]

Ziel von NLP ist üblicherweise die Extraktion von Informationen aus einem Text durch kleinschrittige Textanalyse. Diese Informationen werden dem Text direkt über sogenannte Annotationen angehängt oder anders abgespeichert, etwa in Ontologien, welche in einem späteren Abschnitt genauer erläutert werden.

Beispiele von Forschung und Entwicklung

Die wissenschaftliche Erforschung von NLP entstand Mitte des 20. Jahrhunderts, als der Linguist Noam Chomsky zeigte, dass sich Merkmale der englischen Sprache formalisieren und somit automatisieren lassen [CHO57].

Frühe Projekte wie der Sprachcomputer ELIZA aus den 1960er Jahren waren somit bereits in der Lage eine natürlichsprachliche Frage als Texteingabe entgegenzunehmen und üblicherweise eine plausible Antwort zu liefern.

Heutzutage gibt es ein Vielzahl von Spracherkennungssoftware auf Smartphones oder Smart-Devices wie ALEXA, welche in der Lage sind gesprochene Kommandos zu erkennen und über das Internet an Server zu leiten, welche in der Lage sind diese Kommandos schnell und effizient zu verarbeiten [HAO14].

Herausforderungen von NLP

Durch die Verbindung von Linguistik und Informatik stellt sich NLP generell der Herausforderung sowohl inhärente Probleme der Sprache, als auch Probleme der Automatisierung lösen zu müssen.

Zu den sprachlichen Problemen gehören Aspekte wie Mehrdeutigkeit, Sarkasmus und Umgangssprache bzw. Redewendungen. Diese lassen sich oft nicht einzig und allein durch den gegebenen Text erklären, sondern beruhen auf Kontext und Situation.

Vor allem bei der Erfassung von Semantik, also dem Inhalt der Texte, können komplexe Probleme auftreten. So gibt es beispielsweise Mehrdeutigkeiten, welche selbst für den Menschen nicht einfach zuzuordnen sind. So kann „*Die Betrachtung des Studenten*“ etwa bedeuten, dass der Student etwas betrachtet. Genauso kann es jedoch sein, dass jemand anderes den Studenten betrachtet.

Eine andere häufige Fehlerquelle ist das korrekte Zuordnen von sprachlichen Bezügen. Betrachtet man die Frage „*Verkaufen Sie Handys und Compu-*

ter von Samsung“, so ist diese eindeutig grammatikalisch korrekt. Dennoch ist es uneindeutig, ob sich die Frage auf nach der Marke der Handys richtet, oder ob dies lediglich auf die Computer bezogen ist.

Noch komplexer wird es, wenn sich diese Bezüge über verschiedenen Sätze bzw. Teilsätze erstrecken. So ist der Satz „*Tom schenkt Tim ein Fahrrad, weil er nett ist*“ ebenfalls auf zwei verschiedenen Arten zu deuten. Einerseits kann es sein, das Tom das Fahrrad verschenkt, weil Tom ein netter Mensch ist, andererseits kann es sein, das Tom das Fahrrad verschenkt, weil Tim nett zu ihm ist.

Weiterhin können auch bei der korrekten Erfassung des Sprachinhalts weitere Probleme die Kommunikation stark beeinflussen. So kann etwa die Intention der Sprache dem Inhalt widersprechen, wie es bei Sarkasmus üblicherweise der Fall ist.

Bei der Implementierung ergeben sich dann Probleme, trotz der sprachlichen Besonderheiten konsistente Regelmäßigkeiten als Grundlage zur Automatisierung zu finden.

Linguistische Analyse

Um einen Text möglichst Fehlerfrei verarbeiten zu können bedarf es also einem kleinschrittigen und präzisen Verfahren. Dieses bezeichnet man als linguistische Analyse.

Sie ist aufgeteilt in vier wichtige Schritte, welche aufeinander aufbauen und an Komplexität zunehmen. Im folgenden sind diese Schritte aufgeführt und kurz erläutert [RS18]:

1. Morphologische Analyse - Die Zerlegung von Wörtern bezüglich ihrer Struktur. Die Komposition aus Präfix, Wortstamm und Suffix von Wörtern wird erkannt, um Fall, Tempus, Numerus etc. des Wortes zu bestimmen. Im Englischen zeigt so die Endung -ed bei Verben die Vergangenheitsform an. Dabei ist zu beachten, dass Mehrdeutigkeiten entstehen können.
2. Syntax - Aufbau von Sätzen durch einzelne Wörter. Jede Sprache besitzt syntaktische Regelungen bezüglich der auftauchenden Wörter; im Deutschen folgt etwa auf einen Artikel irgendwann ein Substantiv oder bestimmte Signalwörter geben Satztyp und Tempus an. Auf Basis dieses Wissens können formale und strukturelle Analysen der Textbestandteile erfolgen.

3. Semantiken - Die Identifikation der Bedeutung von einzelnen Sätzen und Wörtern. Die Semantik eines Textabschnittes wird häufig als „Logik“ bezeichnet und fragt nach dessen Bedeutung bzw. Thema. Semantische Deutungen können anhand von Kompositionen einzelner Wörter und Sätze auf Basis der semantischen Analyse erkannt werden.
4. Kontext - Darstellung von satzübergreifenden Zusammenhängen syntaktischer und semantischer Natur. Mehrere Sätze können über Konstrukte wie etwa Pronomen verbunden sein, wenn sich das Pronomen des einen Satzes auf ein Subjekt des anderen bezieht.

Diese Aufteilung lässt sich jedoch in noch weitere Teilschritte für die Spracherkennung und -generierung als NLP-Architektur unterteilen. Diese wird im folgenden Abschnitt genauer erläutert.

2.1.2 Umsetzung der Verarbeitungsschritte natürlicher Sprache

Dieser Abschnitt beschäftigt sich primär mit den verschiedenen Schritten der sprachlichen Analyse. Diese Differenzierung kann dabei, wie zuvor beschrieben, anhand der linguistischen Analyse unterschiedlicher Bestandteile von Sprache vorgenommen werden. Diese sukzessive Verarbeitung wird im folgenden anhand verschiedener Algorithmen und Verfahren erläutert, die im wesentlichen in [RS18] zu finden sind.

Das Ziel jeder Analyse ist es, Wissen über einen Teilbereich von Sprache zu generieren. Das Wissen über den Zusammenhang von Wörtern eines Textes kann eindeutig sein, jedoch können besonders bei der morphologischen Analyse sprachliche Mehrdeutigkeiten auftreten. Dies wird in 2.1.3-Morphologie näher erläutert.

Die Präzision dieser Aussagen fällt dementsprechend abhängig von der Treffergenauigkeit und dem Abdeckungsgrad der linguistischen Analyse durch die verwendeten Verfahren ab. Es ist dabei zu beachten, dass die Analyse typischerweise als Architektur aufgebaut ist, da die verschiedenen Schritte aufeinander aufbauen [COP04]. Somit können, basierend auf den Ergebnissen, inhaltliche Aussagen und Zusammenhänge über den zugrundeliegenden Text getroffen werden.

[COP04] identifiziert dabei eine allgemeine NLP-Architektur, die basierend auf der linguistischen Analyse diese Teilbereiche voneinander abgrenzt.

Für ein besseres Verständnis der Verfahren werden die für diese Arbeit relevanten Schritte in [RS18] wie folgt kurz erläutert:

1. Input Processing - Erkennung der Dokumentensprache und Normalisierung des Textes. Im ersten Schritt geht es um das korrekte Format des Eingabedokumentes für die Verarbeitung. Dieser Vorgang stellt jedoch noch keine Analyse dar, sondern dient lediglich der Vorbereitung.
2. Morphologische Analyse - Die meisten Sprachen sind im Bezug auf ihre Grammatik und syntaktische Struktur der Wörter in Systemen abgeschlossen, etwa durch Modellierung mittels Automaten zur Erzeugung der Worte. Auch können Vokabeln in Wörterbüchern/Lexika gesammelt und Regeln auf ihnen formuliert und gespeichert werden.
3. Part-of-Speech-Tagging - Die einzelnen Wörter eines Satzes werden im Hinblick auf den Sach- oder Satzzusammenhang, wie auch auf die Stellung im Satz hin analysiert. Basierend auf Kontext und/oder Erfahrungswerten bzw. Heuristiken können die Wörter korrekt erfasst und deren Fall getaggt werden. Subjekte, Prädikate usw. werden identifiziert. Dazu kann eine Wissensbasis mit Trainingsdaten und die Einordnung des Kontextes darin verwendet werden.
4. Parsing - Die Ergebnisse der vorherigen Schritte werden weiter verarbeitet und in ein standardisiertes Format gebracht. Syntaktische Zusammenhänge, wie etwa das Zusammenfassen von Wörtern zu einer gemeinsamen Bedeutungsphrase und das Zuordnen von Verben zu einem Nomen können nach dem Parsing dargestellt werden.
5. Disambiguation - Die Entfernung von Mehrdeutigkeiten auf Basis der Ergebnisse des Parsings stellt einen entscheidenden Schritt für die semantischen Analysen dar. Nur wenn die Aussage und der Kontext eines Satzes klar erkennbar sind, kann dessen Semantik gezielt abgeleitet werden.
6. Context Module - Textbausteine, deren Interpretation semantisch auf dem Kontext anderer Sätze oder Wörtern beruhen, können erfasst werden. Bei Anaphern ist etwa das Verständnis des aktuellen Kontexts abhängig von einer vorherigen Deutung.

7. Text Planning - Die Sachzusammenhänge und Semantiken, die aus dem zugrundeliegenden Text extrahiert wurden, werden für die Darstellung im fertigen Text definiert. Es wird festgelegt, welche der Bedeutungen qualitativ übertragen und vermittelt werden sollen. Die Reihenfolge und Umfang der behandelten Themen wird geschätzt und definiert.
8. Tactical Generation - Die Bedeutungen werden in Form konkreter Zeichenketten generiert, die die gewünschte Bedeutung enthalten. Diese Textbausteine basieren häufig direkt auf den Ergebnissen des vorherigen Parsings, da dort schon Bedeutungen zusammengefasst werden können. Der quantitative Teil des Textes wird erzeugt.
9. Morphological Generation - Die erzeugten Sätze werden morphologisch an die Rahmenbedingungen der verwendeten Sprache angepasst. Grammatiken und Regeln der Satzkonstruktion werden auf die erzeugten Wörter angewendet, sodass ein lesbarer und nachvollziehbarer Text entsteht.
10. Output Processing - Der Text wird nun, nachdem dieser inhaltlich komplett erzeugt vorliegt, an das Design und Format des jeweiligen Einsatzzwecks angepasst. Sollte Text, anders als in dieser Arbeit der Einfachheit halber angenommen, nicht als geschriebenes Wort ohne besondere Formatierung vorliegen, kann dieser einem *Formatter* zur Designanpassung übergeben werden. Ferner kann etwa mithilfe von Text-to-Speech eine Audioausgabe erfolgen.

Wie später gezeigt wird, kann die in dieser Arbeit beschriebene Entwicklung der Software-Prototypen als eine Umsetzung eben dieser NLP-Architektur verstanden werden.

2.1.3 Syntaktische Analyse von Wörtern und Sätzen

In diesem Abschnitt wird die Zerlegung der Syntax anhand logischer Modelle von Sprache beschrieben. Begonnen wird mit der einfachen Struktur und Zusammensetzung einzelner Wörter aus der Grammatik. Die Kombination aus Wortstämmen mit Prä- und Suffixen kann etwa in einem Lexikon gesucht werden, dass alle Wörter einer Sprache enthält. Basierend auf dieser Morphologie eines Wortes kann dann in der Analyse mehrerer Wörter auf die Rolle von Wörtern im Satzzusammenhang geschlossen werden.

Morphologie

Dieser Abschnitt befasst sich genauer mit der Analyse der Struktur und Zusammensetzung von Wörtern. Dies wird als morphologische Analyse bezeichnet. Wie im folgenden beschrieben, können auf dieser Basis logische Modelle wie etwa Automaten erstellt werden, die zulässige Wörter aus einer Sprache auf Basis grammatikalischer Regeln erzeugen können. Für ein besseres Verständnis der im Rahmen dieser Arbeit vorgestellten Ergebnisse beschränkt sich diese Arbeit auf die vergleichsweise einfach strukturierte englische Sprache. Aufgrund der großen Fortschritte und Erfahrungen in der Sprachforschung, werden die folgenden Beispiele auf Englisch behandelt.

In der englischen Sprache besteht jedes Wort aus einem Wortstamm, sowie den sogenannten Affixen. In einem Satz könnte etwa das Wort „*believable*“ auftauchen. Das Wort stammt von der Grundform „(to) *believe*“ (zu Deutsch: „glauben“) ab. Das *-able* stellt ein Suffix dar, also eine Wortendung. Zusammen mit den Präfixen, die dem Wortstamm vorangestellt sind, gehört *-able* somit zur Kategorie der Affixe.

Das Ziel der morphologischen Analyse ist nun, die Zusammensetzung des Wortes nachzuvollziehen und die Information für die weitere Verarbeitung bereitzustellen. Die Generierung des Wortes *believable* kann anhand von Buchstabierregeln erfolgen, die alle zulässigen Wörter auf englisch anhand der Wortstämme und Affixe erzeugen können. Es fällt auf, dass in unserem Beispiel *believable* vor dem Hinzufügen des Suffixes *-able* zunächst das *-e* von der Grundform (to) *believe* entfernt werden muss. Eine Grammatik, die zur Erzeugung dieser Wörter genutzt werden kann, muss dementsprechend eine solche Regel vorhanden sein. Gemäß [RS18] bedeutet dies formal ausgedrückt:

$$e \rightarrow \varepsilon_able$$

In der Logik kann man diese Regeln als nicht-deterministischer *Zustandswandler* visualisieren.² In Abb. 2.1 ist der Zustandswandler gezeigt, die das Suffix *-able* an entsprechende Wörter anhängen kann.

²In [RS18] wurde darauf hingewiesen, dass nicht-deterministische Zustandswandler trotzdem mithilfe der *Potenzmengenkonstruktion* in einen äquivalenten deterministischen Zustandswandler überführt werden kann. Sie kommen somit zu einem eindeutigen Ergebnis *erzeugbar* oder *nicht erzeugbar* in der Sprache.

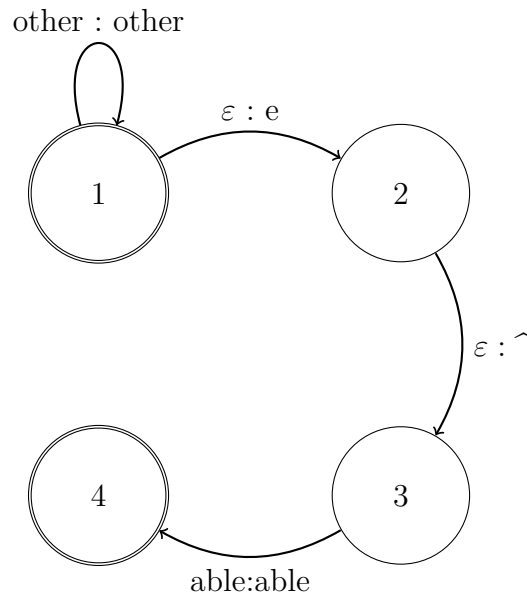


Abbildung 2.1: Nicht-deterministischer Zustandswandler der able-Regel [RS18]

Um alle zulässigen Wörter sammeln und identifizieren zu können, muss die morphologische Analyse über eine Menge an Informationen über die Sprache verfügen. Die Bestandteile von Wörtern können dabei in verschiedenen Lexika enthalten sein, welche die entsprechenden Informationen bereitstellen. Gemäß [COP04] bedarf es dabei drei verschiedener Lexika:

1. Liste von *Affixen* (zusammen mit den Informationen, die damit verbunden sind, z.B. Negation „un-“)
2. Liste von *unregelmäßigen Wörtern* (zusammen mit den Informationen, die damit verbunden sind, z.B. „went“:simple past, „(to) go“)
3. Liste von *Wortstämmen* (zusammen mit syntaktischer Kategorie, z.B. „believe“:verb)

Es ist zu beachten, dass die in der morphologischen Analyse verwendeten Buchstabierregeln zwar zu jeweils eindeutigen Ergebnissen kommen können, die Analyse selbst jedoch nicht eindeutig sein muss. Dies kann, wie bereits zuvor beschreiben, durch sprachliche Mehrdeutigkeiten entstehen.

Das Wort „*even*“ könnte isoliert etwa einmal als *gleichmäßig*, also als Adjektiv kategorisiert, ins Deutsche übersetzt werden, jedoch auch sinngemäß *sogar*, also das Adverb mit einem oder mehreren Bezugswörtern meinen.³ Die mehreren möglichen Bedeutungen des einzelnen Wortes können jedoch in der Regel durch die Position und grammatikalische Funktion im Satz aufgelöst werden. Diese Eigenschaften werden durch das *Part-of-Speech-Tagging* identifiziert.

Part-Of-Speech-Tagging

Die Ergebnisse der morphologischen Analyse von mehreren Wörtern, etwa in einem Satz, können zur Analyse von Zusammenhängen zwischen diesen Wörtern genutzt werden. Somit kann die grammatikalische Struktur des Satzes auf Basis der einzelnen Wörter dargestellt werden. Der grammatikalische Typ jedes Wortes wird durch das Part-Of-Speech-Tagging festgelegt und jedes Wort wird *annotiert*. Wie später in dieser Arbeit beschrieben wird, dienen solche Annotationen in NLP-Werkzeugen der Sammlung der bekannten Informationen zu einem Wort oder Satz.

Zur Festlegung dieser grammatikalischen Funktionen werden sog. *POS-Tagger* eingesetzt, die verschiedene Worttypen als Kategorien beinhalten. Diese werden dann jedem Wort eines Satzes als Annotation hinzugefügt. Die Aufgabe ist es nun, die korrekte Funktion eines Wortes im Satz auf Basis der morphologischen Analyse zu identifizieren.

Als Beispiel soll der Satz

I love trains.

dienen. Das Part-of-Speech tagging zum Wort „*trains*“ soll hier veranschaulicht werden. Das Wort „*trains*“ allein ist ein Nomen im Plural, jedoch könnte es sowohl das Subjekt, als auch das Objekt im Satz darstellen. Im Englischen sollte ein POS-Tagger aber erkennen, dass *Subjekt-Prädikat-Objekt* eine gültige Satzkonstruktion darstellt, die zu dem Beispielsatz passt. Dementsprechend entfällt die Möglichkeit von *trains* als Subjekt, da es hinter einem Verb als Akkusativobjekt des Satzes auftaucht.

³Aus Gründen der Vereinfachung wurde sich hier auf zwei erkennbar verschiedene Bedeutungen beschränkt. Tatsächlich stellt sich neben Bestimmung des grammatikalischen Typs auch die Frage, ob sich die verschiedenen Übersetzungen derselben Kategorie, z.B. *even* als *gerade* und *gleichmäßig*, in ihrer Bedeutung im Sachzusammenhang unterscheiden.

Es existieren verschiedene Techniken für diese Beurteilungen der POS-Tagger. Verschiedene gültige Satzkonstruktionen sind häufig als sog. *corpora* gespeichert. Diese stellen repräsentative Trainingsdaten dar, die beispielsweise aus der Zeitung *Wall-Street-Journal* stammen. In diesem Corpus könnten etwa die annotierten Sätze

People(subj) love(vb) trains(obj) .(punc) und
I(subj) go(vb) home(obj) .(punc)

enthalten sein.⁴

Mithilfe dieser Menge von Daten kann POS-Tagging auch heuristisch, d.h. auf Basis von Wahrscheinlichkeiten bzw. Metriken, erfolgen, indem das Auftauchen verschiedener Satzbauteile im Trainingscorpus gezählt wird. Auf dieser Basis können dann Wahrscheinlichkeiten für verschiedene Möglichkeiten berechnet werden, wenn es mehrere mögliche Zuordnungen gibt.

Nutzen syntaktischer Analyse für NLP

Ein wichtiges Einsatzgebiet für POS-Tagger ist auch die Wortvorhersage bzw. *prediction* auf Basis solcher Heuristiken. Diese kann im Kontext der zuvor vorgenommen Annotationen von Wörtern erfolgen. Beispielsweise werden sog. *n-gramme* verwendet, um unter Berücksichtigung der Annotation von *n* gelesenen Wörtern die Möglichkeiten für das folgende Wort zu berechnen. In [RS18] findet sich ein vollständiges Beispiel anhand eines *Bigramms* mit Corpus.

Die heutigen Verfahren zur syntaktischen Analyse nutzen zur Steigerung der Genauigkeit u.a. Verfahren aus dem maschinellen und automatisierten Lernen und können somit hohe Trefferquoten erzielen.⁵ Um ein tieferes Verständnis von Texten zu erzielen, reicht die in diesem Abschnitt beschriebene syntaktische Analyse anhand der Grammatik nicht aus. Die tatsächliche

⁴Die tatsächliche Annotation kann durch POS-Tagger i.d.R. feiner erfolgen, sodass etwa Pronomen von Subjekten unterschieden werden. Auch können, wie später in dieser Arbeit gezeigt wird, eigene Annotationen bei Bedarf nach neuen Kategorien hinzugefügt werden. Der Corpus hier stellt ein stark vereinfachtes Beispiel dar. Ein ausführliches Beispiel unter der Verwendung der CLAWS-7-POS-Taggers findet sich in [RS18].

⁵In [SHE07] wird ein POS-Tagger auf der Basis von bidirektionaler Analyse und Machine-Learning zum Erstellen der Trainingsdaten beschrieben. Obwohl nur auf Basis der Textsyntax erzielt, sind die Ergebnisse zu 97,33% korrekt.

Abbildung von Wissen über semantische Zusammenhänge stellt eine verlässlichere Möglichkeit dar, sprachliche Zusammenhänge zu identifizieren.

2.1.4 Umgang mit Semantik

Wie im letzten Abschnitt aufgezeigt wurde, können nicht alle Informationen aus einem Text eindeutig anhand der Erfassung der Grammatik eines Satzes extrahiert werden. Die Syntax eines Satzes ist häufig mehrdeutig, da der Zusammenhang einzelner Wörter verschieden interpretiert werden kann. Auf der semantischen Ebene, d.h. der Bedeutung, existieren jedoch bereits bei einzelnen Wörtern verschiedene Deutungen.

Schon bei der Mensch-Mensch-Kommunikation kommt es somit häufig zu Missverständnisse, wenn die Bedeutung hinter einer Aussage nicht eindeutig ist. Auf Basis des Kontextes einer Aussage bzw. der Domäne des Inhalts kann jedoch meist zu einer mehrdeutigen Aussage eine eindeutige Deutung identifiziert werden.

NLP beschäftigt sich daher auch intensiv mit der Abbildung und Modellierung der Semantik von Sprache, die in der Logik häufig als „*Logik einer Aussage*“ bezeichnet wird.⁶ Die Formalisierung mehrerer Bedeutungen kann dabei mithilfe von Symbolen der Aussagenlogik erfolgen, die Symbole dienen dabei als Platzhalter für Wörter bzw. Satzteile. Etwa das Beispiel aus [RS18]:

Every man loves a woman.

Zu deutsch „*Jeder Mann liebt...*“ ist semantisch mehrdeutig zwischen

1. „...*die eine Frau*“: $\exists y[\text{woman}'(y) \wedge \forall x[\text{man}'(x) \Rightarrow \text{love}'(x, y)]]$ und
2. „...*irgendeine Frau*“: $\forall x[\text{man}'(x) \Rightarrow \exists y[\text{woman}'(y) \wedge \text{love}'(x, y)]]$

Bereits durch den Menschen fällt die Deutung anhand von Symbolen der Aussagenlogik nicht leicht. Daher wurden verschiedene Arten von Zusammenhängen, u.a. auf Basis von Symbolen der Mengenlehre, anhand welcher die Deutung vereinfacht und klarer erkennbar wird.

Meaning-Postulates

Linguisten und Philosophen haben Regeln für die Abbildung von Zusammenhängen auf eindeutige Kategorien formalisiert. Dabei wurden verschiedene

⁶Bereits in den 1980er Jahren wurden Konzepte auf Basis logischer Modelle, mit denen Semantik abgebildet werden kann, mithilfe von Theorembeweisern und Model-Checking beschrieben. [SB88] [KN85]

Bedeutungen von Wortzusammenhängen identifiziert, die als (Bedeutungs-)Mengenbeziehung bzw. Relation zwischen Symbolen visualisiert werden können [CAR52]:

1. *Meaning-Postulates*: A beinhaltet B und C

Der Begriff A setzt sich logisch aus der Relation zwischen B und C zusammen, etwa:

$$\forall x[\text{bachelor}(x) \Rightarrow \text{man}(x) \wedge \text{unmarried}(x)]$$

sinngemäß: „Ein Junggeselle ist ein Mann und nicht verheiratet.“

2. *Hyponymy*: A ist ein B

Ein Hund ist ein Tier.

Also hat ein Hund auch alle Eigenschaften eines Tieres. Auf Basis solchen Wissens kann eine Kategorisierung erfolgen bzw. lassen sich verschiedene Ausprägungen der gleichen Sache zusammenfassen [RS18].

3. *Meronymy*: $A \subseteq B$, A ist Teil von B

4. *Synonymy*: $A \equiv B$, A ist semantisch äquivalent zu B

5. *Antonymy*: $A \equiv \neg B$, A bedeutet das Gegenteil von B

Feinheiten der Deutung, wie etwa der Unterschied von „A semantisch äquivalent zu B“ zu „A ist ein B“ müssen dabei während der Erstellung semantischer Modelle unterschieden werden, da die Kategorisierung sonst zu grob ausfällt. Es können sonst falsche Zusammenhänge und Beziehungen entstehen, die Fehlinterpretationen der Aussagen erzeugen.

In der Praxis stellt sich daher die Frage, in welchem Umfang Semantiken für NLP tatsächlich für eine Verbesserung der Textanalyse sorgen und wie detailliert eine Modellierung dazu erfolgen muss. Mit diesem Ansatz beschäftigen sich die *Ontologien* in Kapitel 2.2.

2.2 Ontologien

Dieser Abschnitt befasst sich genauer mit dem Begriff der Ontologien im Kontext der Informatik.

2.2.1 Definition

Ontologien lassen sich allgemein definieren als „formale, schematische Abbildungen eines Wissensbereichs, bestehend aus einem Vokabular und Regeln zu seiner Zusammensetzung“[WE13].

Zunächst bedeutet dies, dass eine Ontologie stets auf nur Wissen aus einer bestimmten Domäne bezieht und nicht versucht das gesamte über die Welt bzw. sehr allgemeines Wissen abzubilden.

Des weiteren Besteht eine Unterteilung in ein Vokabular und in eine Menge von Regeln.

Einer der Aspekte von Ontologien, welche sie für die Informatik so geeignet machen ist, dass sie in maschinenlesbaren Ontologiesprachen verfasst werden. Somit können Computersysteme diese automatisch auslesen und ggf. interpretieren und daraus schlussfolgern.

Da Terminologie und bestimmte Formalitäten zwischen Ontologiesprachen variieren können, wird im Folgenden der Bezug auf die weit verbreitete Ontologiesprache „Web Ontology Language“ (OWL) angenommen.

2.2.2 Aufbau

Der Aufbau von Ontologien lässt sich grob in zwei Aspekte unterteilen:

1. Das Vokabular: Hier sind alle Dinge innerhalb des Wissensbereiches aufgeführt, welcher von der Ontologie abgedeckt werden soll. Dazu gehören sowohl die Oberbegriffe dieser Dinge, als auch die Instanzen selbst.
2. Die Regeln: Diese beschreiben die Dinge genauer und bilden deren Beziehung untereinander ab. Sie können eigenständig existieren oder im Bezug zu den vorhandenen Dingen stehen.

Klassenhierarchie

Die Hierarchie einer Ontologie, also der strukturelle Aufbau, lässt sich in die folgenden drei Komponenten unterteilen:

1. Begriffe: Die Grundlage für die Struktur einer Ontologie stellt der hierarchische Aufbau von *Begriffen* (*concepts*). Diese Repräsentieren eine Art Klassenstruktur und werden daher oft auch als Klassen bezeichnet. Sie können in aus der Informatik bekannten Ansätzen wie Ober-

und Unterklassen angeordnet werden. Begriffe sind üblicherweise sehr generell gehalten und könnten etwa die Klasse *Auto* abbilden.

2. Typen: In vielen Fällen lässt sich ein Begriff in verschiedene *Typen* aufteilen. Ein Auto kann beispielsweise nach verschiedenen Automarken typisiert werden. Diese Typen werden dann ebenfalls durch Klassen verkörpert und bilden in dem hierarchischen Aufbau eine Unterklasse des jeweiligen Begriffs. Da sowohl Begriffe, als auch Typen in Klassen verkörpert werden, wird in der Praxis oftmals nicht zwischen den beiden unterschieden. Man kann jedoch üblicherweise über einer Unterklasse als Typ der Oberklasse sprechen.
3. Instanzen: Die sogenannten *Instanzen* (*instances*) verkörpern konkrete Objekte dieser Klassen und werden nicht weiter aufgespalten bzw. typisiert. Ein Instanz wäre dann etwa ein bestimmtes Modell, einer bestimmten Marke von Autos.

Diese Hierarchie ist ggf. beliebig erweiterbar. So könnte man für eine bessere Klassifizierung z.B. die Marke eines Autos zunächst in eine Produktreihe oder nach einer Produkteigenschaft wie *Combi* typisieren und erst davon abhängig die verschiedenen Instanzen zuordnen.

In Abbildung X.X ist eine stark vereinfachte Klassenhierarchie aus Oberklassen, Unterklassen und Instanzen beispielhaft dargestellt.

HIER KLASSENHIERARCHIEBILD EINFÜGEN

Auf der obersten Ebene der Klassen-Hierarchie befindet sich die Klasse *owl:Thing*. Von dieser erben alle Begriffe durch die *hasSubclass*-Relation. Dies ist vergleichbar mit der objektorientierten Programmiersprache *Java* bei welcher alle Klassen direkt oder indirekt von der Oberklasse *Object* erben.

Auf der nächsten Ebene befinden sich die beiden Begriffe *Automobil* und *Person*. Diese können entweder direkt in Instanzen übergehen, oder zunächst durch Typisierung unterteilt werden.

Somit finden sich in der nächsten Ebene die drei nach Marke getrennten Automobil-Typen *VW*, *BMW* und *Audi*. Außerdem werden Autofahrer und Ingenieure als Personen unterschieden. Es ist anzumerken, dass es nicht ausgeschlossen ist, dass ein Ingenieur auch Autofahren ist. Wegen der möglichen Mehrfachvererbung lassen sich jedoch Typen oder Instanzen erzeugen, welche sowohl von Autofahrer, als auch von Ingenieur erben. Somit stellt eine

Unterteilung in Typen ohne gegenseitigen Ausschluss in der Praxis kein Problem dar und kann sogar nützlich sein, wie später anhand von Relationen gezeigt wird.

In der vierten Ebene sind einzig zu Demonstrationszwecken einige der Typen weiter unterteilt und andere nicht. Man sieht, dass die Hierarchie also keineswegs gleichmäßig sein muss, um eine sinnvolle Unterteilung zu gewährleisten.

In der letzten Ebene befinden sich einige Instanzen von Autos welche durch konkrete Modelle dargestellt werden. Theoretisch ließen sich auch diese weiter unterteilen, etwa nach Innenausstattung. Diesbezüglich ist je nach Verwendungszweck der Ontologie genau festzulegen, wie die Hierarchie zu strukturieren und wie weit sie auszuführen ist. Beispielsweise macht es für einen großen Autohersteller keinen Sinn, wenn in einer konzernübergreifenden Ontologie jeder Ingenieur instantiiert ist.

Regeln

Die Regeln in einer Ontologie lassen sich in die folgenden zwei Komponenten unterteilen:

1. Relationen: Einen wichtigen Bestandteil von Ontologien bilden die sogenannten *Relationen (properties)*. Relationen beschreiben Beziehungen zwischen Begriffen und werden oft auch als Eigenschaften verwendet und bezeichnet. Ähnlich zu vielen Programmiersprachen können diese Relationen und Eigenschaften eines Begriffs an die Instanzen des Begriffs vererbt werden. Grundsätzlich ist dabei auch eine Mehrfachvererbung möglich.
2. Axiome: Des Weiteren können Ontologien durch sogenannte *Axiome (restrictions)* genauer modelliert bzw. eingeschränkt werden. Axiome verkörpern dabei Regeln in der Ontologie welche stets zutreffen müssen. Üblicherweise werden diese dazu verwendet, Einschränkungen vorzunehmen, die nicht effizient durch den abgebildeten Wissensbereich darzustellen sind. Dies könnte Beispielsweise in einer Ontologie über verschiedene Automodelle verschiedenster Marken eine generelle Definition von Automobilen sein. So ließe sich die Voraussetzung, dass ein Automobil motorisiert sein muss durch eine Klasse *Bauteil* mit der Unterklasse *Motor* darstellen, welche dann über eine *muss verbaut sein in*-Relation mit der Klasse *Automobil* verbunden ist. Ist es jedoch nicht

Zweck der Ontologie Bauteile genauer abzudecken, wäre dies unnötiger Aufwand, der außerdem die Übersichtlichkeit der Ontologie einschränken könnte.

Abbildung X.X zeigt eine beispielhafte, stark vereinfacht Ontologie zu Automobilen und bestimmten im Kontext relevanten Personen.

HIER RELATIONEN BILD

Die Relationen zwischen den Konten der Ontologie sind dargestellt durch Linien mit einem mittigen Pfeil. Die durchgezogenen Linien zeigen den Grundaufbau der Hierarchie. Die blauen Linien verkörpern die *has subclass*-Beziehung, welche auf Vererbung hinweist. Die violetten Linien stehen für die *has individual*-Beziehung und zeigen auf eine Instanz der Klasse. Außerdem wurden zwei Objekt-Attribute (*object properties*) manuell hinzugefügt. Diese können verwendet werden um Relationen zwischen Klassen aufzuzeigen, welche nicht voneinander erben, sondern sich in unterschiedlichen „Teilbäumen“ der Ontologie befinden. So verweist, dargestellt durch den braunen Pfeil eine *benutzt*-Relation von dem Autofahrer auf das Automobil, bedeutet also das Autofahrer Automobile verwenden. Ähnlich dazu zeigt von der Klasse Ingenieur eine *entwickelt*-Beziehung auf Automobil.

NACHSCHAUEN WEGEN VERERBUNG AUF INSTANZEN.

Verwendung von Ontologien

Wie eingehend bereits erwähnt sind Ontologien vor allem in der Informatik durch ihre Maschinenlesbarkeit wertvoll. Sie sind außerdem auch geeignet komplexere Sachzusammenhänge abzubilden, was mit primitiven Datentypen wie man sie in einer relationalen Datenbank abspeichern kann nicht möglich wäre. Simple Verbindungen wie sie etwa in den in 2.1.4 beschriebenen Bedeutungspostulaten beschrieben sind lassen sich in Ontologien schnell und effizient darstellen.

Betrachte man zum Beispiel das Satzschema „B ist ein A“ so würde dies durch eine einfache *hasSubclass*-Beziehung von der Klasse A zu der Klasse B dargestellt. Erstellt man nun die beiden Instanzen *b1* und *b2* der Klasse B, so liegen implizit bereits die Informationen vor, dass sie beide auch ein A sind. Außerdem ist ihre Homogenität durch die Abstammung von der gleichen Klasse gewährleistet. Daher brauchen die Instanzen hierfür weder eine

Beziehung untereinander, noch zu der Oberklasse A ihrer Elternklasse B .

Grafik mit A,B,b1 und b2 und Tier, Vierbeiner, Hund und Katze

Im Zusammenhang dieser Arbeit sollen später Ontologien dazu verwendet werden den Inhalt betrieblichen Anforderungen aus einem Pflichtenheft abzuspeichern. Pflichtenhefte sind dabei stark domänenbezogen und somit optimal geeignet um in Ontologien abgespeichert zu werden. Des Weiteren lassen sich stets Relationen zwischen den einzelnen Teilbereichen ziehen, was für einen einzelnen Domänenexperten schwierig wäre. Eine üblicherweise vorausgesetzte Eigenschaft für Anforderungen ist Atomarität [QUELLE], was bedeutet, dass jeder Satz nur genau eine Aussage beinhalten soll. Dies hat zur Folge, dass viele betriebliche Anforderungen tatsächlich den vorgestellten, knappen Satzchemata entsprechen.

2.3 Verwandte Arbeiten

Dieses Kapitel stellt einen Überblick über bereits veröffentlichte Arbeiten dar, die sich mit dem Themen Natural Language Processing für Anforderungsmanagement, sowie die Unterstützung durch Werkzeuge beim Requirements-Engineering befassen.

Grundlage dieser Arbeit

K. Zichler und S. Helke stellen in [ZH17] und [?] die wesentlichen Aspekte für den Entwurf einer Softwareunterstützung für Requirements-Engineering in Firmen vor.

In [ZH17] wird zunächst der Bedarf nach solcher Unterstützung besonders für die Automobilindustrie beschrieben, da die Verbesserung herkömmlicher Methoden im Requirements-Engineering dort wesentliches Potential bietet. Der Vorschlag beinhaltet die Unterstützung von Projektteams durch ein Expertenteam aus Requirements-Experten, die Anforderungen aufbereiten und deren konsistente Umsetzung verbessern.

Auf dieser Basis stellen sie in [?] konkret einen Entwurf eines Werkzeugs für die automatisierte Analyse von Projektlastenheften vor. Das Werkzeug gliedert sich dabei in eine Komponente zur Überführung natürlichsprachli-

cher Anforderungen in vordefinierte Sprachschablonen, genannt Requirements-to-Boilerplate-Converter (*R2BC*). Auf dieser Basis arbeitet dann der *Delta-Analyser* der die Unterschiede zwischen mehreren Lastenheften, die in Schablonen überführt sind, darstellen kann. Dabei kommen auch Ontologien und NLP-Methoden zum Einsatz. Die Prototypen basieren dabei auf der Open-Source-Software *GATE* zur sprachlichen Auswertung und sollen als 3-Schichten-Architektur umgesetzt werden.

Diese Arbeiten stellen die wesentliche Grundlage für die in unserer Arbeit entworfenen Prototypen dar. Die Implementierung erfolgte in Zusammenarbeit mit dem Autoren Konstantin Zichler bei der Firma HELLA, die den konkreten Anwendungsfall für seine Entwürfe und unsere Umsetzung dargestellt hat.

Es ist anzumerken, dass eine Vielzahl betrachteter Arbeiten das Verbessern von Requirements mithilfe von Schablonen zum Thema haben, sich diese jedoch auf eine Unterstützung bei der Formulierung von Anforderungen beschränken. Eine Umsetzung einer automatisierten Verbesserung und Anpassung bestehender Anforderungen wird häufig zwar umrissen, jedoch selten implementiert bzw. real getestet.

Literatur

In [KR93] untersuchte Kevin Ryan bereits vor modernen Entwicklungen im Bereich NLP in die Möglichkeiten einer Unterstützung von Requirements-Engineering durch natürlichsprachliche Analyse (NLP). Dabei weist er auf die Vorteile einer formalen Einschränkung bereits bei der Spezifikation von Anforderungen hin, womit die Möglichkeit einer Fehlinterpretation vermindert wird. Ryan stuft jedoch auch die Verifikation von Anforderungen durch NLP als geeignet ein, um Konsistenz und Eindeutigkeit der Anforderungen zu gewährleisten. Dazu generieren sie Schemas aus der Analyse der Anforderungen, die dann inhaltlich mit neu dazukommenden Anforderungen im späteren Entwicklungsprozess abgeglichen werden können. Auch schlägt er die Einteilung besonders von Text im Anforderungsdokument in Kategorien vor, damit die Sichtbarkeit von Anforderungen verbessert wird.

In [FF11] beschäftigen sich Farfelder et. al mit der Unterstützung von Requirements-Experten bei der Anforderungsspezifikation durch vorgefertigte Sprachschablonen und Ontologien, womit Fehler bei der Konsistenz von

Anforderungen vermeiden werden. Sie stellen das Tool *DODT* vor, dass als Alternative zur natürlichsprachlichen Formulierung das Erstellen von Anforderungen anhand von Schablonen anbietet, aus denen der Nutzer im Programm auswählen kann. Dabei greift das Werkzeug auf eine Ontologie zu, welche alle möglichen Begriffe zum Einsetzen in die Schablone enthält. Die Eintragungen des Nutzers stellen dann eine Anforderung dar, die in der Ontologie gespeichert wird.

In [FH14] wird das Vorgehen für die Dokumentation und logische Einordnung von funktionalen Anforderungen mithilfe von Sprachschablonen beschrieben. Dazu werden aus den Anforderungen für jedes Gesamtsystem zunächst die einzelnen Komponenten benannt und einzeln anhand ihrer Eigenschaften gespeichert. Die einzelnen Komponenten des Modells können dann in Beziehung zueinander gesetzt werden. Auf Basis der einzelnen Komponenten und der Beziehungen kann dann ein Netz aus allen Komponenten des Gesamtsystems erstellt werden. Zur Erstellung dieser formalen Modelle mittels verschiedener Sprachschablonen haben sie das Werkzeug *ReqPat* entwickelt, das als Erweiterung in Anforderungsmanagement-Programmen wie *IBM Rational-DOORS* integriert werden kann. Ferner kann zur grafischen Darstellung des erstellten Modells ein Export aus Reqpat in UML-Diagramme erfolgen.

[CA15] beinhaltet einen Vorschlag für das automatisierte Überführen von natürlichsprachlichen Anforderungen in Sprachschablonen mittels NLP. Sie stellen eine Methode vor, die sprachliche Muster in den Anforderungen erkennt und daraus eine Grammatik in Backus-Naur-Form (BNF) erzeugt. Diese Grammatik kann somit zulässige Sätze in einem Anforderungsdokument anhand der Satzbausteine erzeugen bzw. finden. Durch die Übersetzung in eine JAPE-Regel kann auf dieser Basis die Erkennung der gefundenen sprachlichen Regeln durch das NLP-Werkzeug GATE erfolgen, das die Konformität der Anforderungen auf Basis der Regeln prüft.

Heßeler et. al untersuchen in [HE13] formale Prozesse für das Anforderungsmanagement in Unternehmen. Sie zeigen die möglichen Verbesserungen des projektbezogenen Requirements-Engineerings durch die Definition verschiedener Rollen für die beteiligten Entwickler auf. Dabei beschreiben sie explizit Kriterien für die Auswahl von Tools für eine Werkzeugunterstützung der Requirements-Engineers in Firmen. Konkret beschreiben sie Werkzeuge

für die Unterstützung der Formulierung von Anforderungen bzw. für die Verifikation von Anforderungen mithilfe grafischer Modelle. Das Tool *DOORS Telematic* dient dabei zur Sammlung von Anforderungen und der Verknüpfung durch Relationen in einer Datenbank, die alle Anforderungen hierarchisch enthält. Tau Generation2 wird für die Visualisierung von Systemeigenschaften in UML2.0-Diagrammen und für den Export als DOORS-Objekte in die Anforderungs-Datenbank genutzt.

Kapitel 3

Betriebliches Umfeld - Use-Case HELLA

Im betrieblichen Umfeld liegen zu Beginn jeden Entwicklungsprojektes für neue Produkte die Aufgaben und Ziele für die Entwicklung als Dokumente vor. Forschungsergebnisse finden Anwendung in der Vorentwicklungsphase, in der die Eignung der Erkenntnisse für neue Produkte eines Unternehmens evaluiert wird. Die Produktentwicklung unterliegt dabei bestimmten Kriterien und Faktoren, die den unternehmerischen Erfolg beeinflussen. Neben betriebswirtschaftlichen Einflüssen wie der Einordnung des Produktes in der Wertschöpfungskette sind es dabei besonders technische Anforderungen an das Produkt, die definiert und während der Produktentwicklung eingehalten werden müssen. Verschiedenste Akteure aus einem Unternehmen sind dabei an der Festlegung der Anforderungen an ein Entwicklungsprojekt bzw. Produkt beteiligt.

In der Automobilindustrie betrifft dieser Ablauf zumeist die Entwicklung neuer Fahrzeugkomponenten, heutzutage meist elektronische und mechanische Bausteine. Diese Bausteine werden dabei nicht sämtlich vom Fahrzeughersteller (OEM) selbst, sondern durch eine Vielzahl von Zulieferern produziert und entwickelt. Die Produktspezifikationen liegen meist digital als Texte, Tabellen und Grafiken vor und werden an den Zulieferer übermittelt. Nach dem Entwicklungsprozess steht dann die (Serien-)entwicklung und -fertigung des Produktes für das Ausrollen in großen Stückzahlen an den Hersteller, der das zugelieferte Produkt dann in seinen Produkten verwendet. Um dies zu erreichen, müssen während des gesamten Prozesses die Anforderungen, die das Systemumfeld des Fahrzeugherstellers hat, berücksichtigt und

eingehalten werden.

Die Anforderungen an das Produkt, etwa technische Rahmenbedingungen, werden dabei von vielen verschiedenen Domänenexperten beim OEM formuliert und in das sogenannte Pflichtenheft für die Entwicklung eingetragen. Beteiligte sind etwa Produktdesigner, Ingenieure und Systemtechniker, die an verschiedenen Stellen im Lastenheft Anforderungen an eine Komponente festlegen. Diese Beteiligten sind in der Regel auf ihren Bereich spezialisiert und nicht interdisziplinär, zudem gibt es sprachliche Eigenheiten der Autoren und unternehmensinterne Richtlinien für die Formulierung, die das Verständnis erschweren können. Demzufolge sammeln sich im Lastenheft verschiedenste Merkmale einer Komponente, die aber nicht im Bezug zueinander stehen und sich im schlimmsten Fall gegenseitig ausschließen.

Durch diese fachliche Breite und Tiefe der Spezifikationen im Pflichtenheft, aber auch durch den Umfang des Lastenheftes von mehreren tausend Seiten, kommt es häufig insbesondere zu Verständnisproblemen auf Seite des Zulieferers. Die Gewichtung einzelner Anforderungen in einem größeren Systemkontext fällt dort schwer, da nun Projektteammitglieder, die an der Entstehung des Lastenheftes nicht beteiligt waren, dieses verstehen und ein Produkt entwickeln sollen, dass möglichst alle Anforderungen berücksichtigt. In Texten muss also nach Zusammenhängen und Bezügen zwischen mehreren Anforderungen gesucht werden, damit die Korrektheit des späteren Produktes gewährleistet ist.

Die Analyse von Zusammenhängen zwischen Anforderungen stellt dabei aus Gründen der Effizienz ein Problem dar, wenn jeder Beteiligte von Hand die für ihn relevanten Anforderungen aus dem Lastenheft extrahieren muss. Auch müssen die Lastenhefte an die Formulierungen und Ausdrucksweisen für Requirements-Management im Unternehmen angepasst werden. Bislang gibt es jedoch kaum Werkzeugunterstützung, die effiziente Möglichkeiten zur automatisierten Überarbeitung und Anpassung einzelner Anforderungen aus dem Dokument bietet. Ansätze aus dem *Natural-Language-Processing* (NLP) stellen gleichzeitig vielversprechende Forschungsfelder in der Informatik dar, die eine solche automatisierte Verarbeitung auf Basis von Sprachanalyse ermöglichen. Syntax und Semantik der einzelnen Sätze und Zusammenhänge in Texten können auf Basis aktueller Trends wie Machine-Learning und dynamischer Programmierung zunehmend besser abgebildet werden.

3.1 Allgemeines zu HELLA

In diesem Abschnitt wird zur besseren Einordnung die Firma HELLA kurz beschrieben. Dies beinhaltet die Geschichte von HELLA und die momentane Aufstellung des Unternehmens. [HE18] [HE19]

Geschichte

Die HELLA GmbH & Co. KGaA wurde 1899 von Sally Windmüller unter dem damaligen Namen „Westfälische Metall-Industrie Aktien- Gesellschaft(WMI)“ in Lippstadt gegründet, wo sich bis heute auch ihr Hauptsitz befindet.

Die Hergestellten Produkte beschränkten sich zu diesem Zeitpunkt auf Ballhupen und Lampen für Kutschen. Im Jahr 1908 tauchte mit dem ersten entwickelten elektrischen Scheinwerfer auch der Name *HELLA* zum ersten mal als Warenzeichen auf. Dieser wurde jedoch erst 1986 offiziell in die Firmenbeschreibung aufgenommen. Es wird davon ausgegangen, dass der Name eine Mischung aus der Anlehnung an den Namen von Windmüllers Frau Helene und dem Wortspiel mit „heller“ ist.

Die erste Tochtergesellschaft gründete HELLA 1951 in Todtnau und dem Namen „Metallwerke Todtnau“. 1961 begann das Unternehmen sich auch international aufzustellen. Dies geschah durch die Gründung der ersten internationalen Produktionsstätte in Mentone in Australien. Das Zentrallager, welches auch heute noch unter dem Namen *HELLA Distribution GMBH* besteht, wurde 1973 in dem Lippstadt nahegelegenen Erwitte errichtet.

Aktueller Stand des Unternehmens

Die HELLA ist in die vier Geschäftsbereiche Elektronik, Scheinwerfer, *After-market* und *Special Applications* unterteilt und verfügt über 125 Standorte in mehr als 35 Ländern.

Nach dem Geschäftsjahr 2017/2018 werden insgesamt über 40.000 Mitarbeiter beschäftigt. Davon arbeiten mehr als 7000 Mitarbeiter in der Forschung und Entwicklung, wodurch sich das Unternehmen mehr Sicherheit und Wachstum für die Zukunft verspricht.

So konnte HELLA in diesem Geschäftsjahr fast 7,1 Milliarden Euro Umsatz verbuchen, wovon ca. ein Drittel außerhalb des europäischen Marktes erzielt wurde. Damit gehört HELLA international zu den Top 40 der welt-

weiten Automobilzulieferer und befindet sich unter den Top 100 der größten deutschen Industrieunternehmen.

3.2 Requirements-Engineering innerhalb der Firma

In diesem Abschnitt wird das Requirements-Engineering (RE) der Firma HELLA genauer behandelt. Dies beinhaltet das Vorstellen der zuständigen Abteilungen, sowie die genaue Darstellung der HELLA-internen RE-Prozesse. Des Weiteren werden die bestehenden Probleme dieser Prozesse anhand von Use-Cases (Anwendungsfällen) aufgeführt. In den darauffolgenden Abschnitten werden daraus Anforderungen an eine Lösung erhoben, welche mit den Konzepten der Werkzeuge umzusetzen versucht wird.

Die getroffenen Aussagen stützen sich dabei auf die angegebenen Lektüren, Unterlagen der Firma und/oder Aussagen Firmen-interner Experten.

3.2.1 Abteilung für RE

Die in der HELLA überwiegend zuständige Abteilung für Requirements ist die *Processes, Methods and Tools*-Abteilung (PMT). Wörtlich übersetzt bedeutet dies *Prozesse, Methoden und Werkzeuge*. Die Abteilung ist unter anderem für:

1. die Standardisierung von Prozessen und das reibungsfrei Verknüpfen von Arbeitsabläufen,
2. die Einführung von neuen Standards und die Hilfe bei der Umsetzung dieser,
3. die Entwicklung neuer und effizienterer Arbeitsmethoden zur Einsparung von Zeit und Kosten,
4. die Entwicklung, Einführung und Verwaltung von neuen (Software-) Werkzeugen, welche in der Lage sind einen Mehrwert für die Firma zu erzeugen,
5. das Empfangen Firmen-externer (Anforderungs-) Dokumente, welche vor dem Weiterleiten gegebenenfalls zunächst in die Firmen-internen Standards angepasst werden müssen,

6. die Unterstützung von Mitarbeitern und Projektteams bei dem Umgang und der Verarbeitung von Projektlastenheften

Oftmals steht diese Abteilung in Verbindung mit den Abteilungen der Vorentwicklung. Dabei spielt vor allem die System-Sparte der elektronischen Vorentwicklung (Electronics - Advanced Engineering Systems, E-AES) eine wichtige Rolle. Hier versuchen Mitarbeiter mit Hilfe von Lastenheften einen kompletten Überblick eines Systems zu erlangen um darauf bezogene Projekte besser einschätzen und durchführen zu können.

Was zunächst wie eine recht simple Aufgabenstellung erscheint ist tatsächlich ein hochkomplexer und fehleranfälliger Vorgang. Laut der Aussagen von internen Experten, gibt es nur wenige, die dazu überhaupt in der Lage sein, da dies viele verschiedene Qualitäten eines Mitarbeiters erfordert. Hierzu gehören unter anderem eine hohes Maß an Fachkenntnis für alle relevanten Bereiche des Produkts, ein gründliches und somit aufwendiges Erfassen aller Aspekte des Produkts und die Fähigkeit mit Hilfe dieses Wissens mögliche Probleme oder Fehlerquellen des Produkts zu erfassen.

Diese Aufgabe kann jedoch von Mitarbeitern der *PMT*-Abteilung (wie in 6. erwähnt) unterstützt werden. Da es erforderlich ist, dass ein einzelner Mitarbeiter der *E-AES* einen Überblick erhält, bietet die klassische Arbeitsteilung, bei welcher der möglicherweise enorme Umfang dieser Projektlastenhefte auf verschiedene Mitarbeiter aufgeteilt wird, keine Lösung.

Dennoch können klar formulierte, gut sortierte und strukturierte Anforderungen die Arbeit des *E-AES*-Mitarbeiters stark vereinfachen. Dies hilft zwar auch ein wenig bei dem benötigten Arbeitsaufwand des einzelnen Mitarbeiters, verringert aber vor allem die Komplexität und Fehleranfälligkeit des Prozesses. Des weiteren kann die *PMT*-Abteilung bei Rückfragen als Sprachrohr zwischen dem OEM und den internen Mitarbeitern fungieren, da die unkontrollierte Weitergabe von Informationen zwischen einzelnen Mitarbeitern in vielen Konzernen problematisch gesehen wird.

3.2.2 RE-Prozess

In diesem Abschnitt wird der RE-Prozess von HELLA genauer dargestellt. Hierzu wird dieser zunächst in den Kontext eines Entwicklungsprojektes für ein neues Produkt eingeordnet. Daraufhin werden die einzelnen Schritte des Requirements-Engineering genauer beleuchtet.

Einordnung in den Entwicklungsprozess

Üblicherweise steht der RE-Prozess direkt zu Beginn jeder Entwicklung. Entweder gibt es ein neues Produktkonzept, zu welchen die Anforderungen abhängig von den geplanten Aufgaben und Eigenschaften des Produktes erhoben werden, oder es wird ein Auftrag eingereicht, zu welchem bereits ein komplettes Anforderungsdokument existiert, welches es dann umzusetzen gilt.

In der Ausgangssituation liegt also ein Dokument vor in welchem alle Anforderungen an das Produkt zusammengefasst sind. Dies bedeutet auch, dass die Entwicklung erfolgreich war, sobald alle dieser Anforderungen durch das Produkt erfüllt werden. Dies muss jedoch auch für die Anforderungen gelten, welche sich innerhalb der Entwicklung ändern bzw. hinzugefügt werden.

Dies verdeutlicht, dass sich der Prozess des Requirements-Engineering nicht auf den Beginn des Gesamtprozesses, beschränken lässt. Dennoch befindet sich dort der wichtigste Bereich des RE, da zumindest in hohem Maße feststehen muss, was es zu entwickeln gilt, bevor mit der Entwicklung begonnen werden kann.

Abbildung X.X zeigt die planmäßige Einordnung von RE in den Gesamtprozess der Produktentwicklung bei HELLA.

HIER DAS V MODELL VOM ENTWICKLUNGSPROZESS BEI HELLA

Hierbei ist anzumerken, dass in dem Modell zwischen den Systemanforderungen (*system requirements*) und den funktionalen Anforderungen (*functional requirements*) unterschieden wird.

Die Systemanforderungen¹ beschreiben Anforderungen, die das System in der Rolle als Umfeld für die Funktionalitäten zu erfüllen hat. Dazu gehören üblicherweise Eckdaten des Systems welche eher generelle Informationen bieten. Eine beispielhafte Anforderung könnte etwa lauten: „Das System muss einer Spannung von bis zu 48V standhalten.“ oder „Das System soll den Standards XY der Norm Z entsprechen.“

Funktionale Anforderungen sind hingegen etwas spezifischer und beschreiben das genaue Handeln einzelner Komponenten und die Ausführung von Funktionalitäten. Hierzu könnten beispielhafte Anforderungen wie folgt aussehen: „Wenn eine Spannung von 24V erreicht wird, soll Komponente XY

¹auch oft als nicht-funktionale Anforderungen bezeichnet

die Funktion *Z* ausführen.“ oder „Die Funktion *Z* lässt die Leuchten *A* und *B* für 10 Sekunden leuchten.“

Die Unterscheidung zwischen den verschiedenen Anforderungstypen kann jedoch sehr problematisch sein, da sich deren Definition in verschiedenen Firmen unterscheiden kann. Da es also keine Einheitliche Definition gibt und wir uns im Rahmen von natürlicher Sprache bewegen, ist es praktisch unmöglich alle Anforderungen eines Dokuments einheitlich und eindeutig den Kategorien zuzuweisen. Dies ist jedoch auch nicht zwangsläufig nötig, wenn man sich dazu entscheidet alle Anforderungen zunächst gleich und dann abhängig von Inhalt zu verarbeiten. Daher verzichten auch die später vorgestellten Konzepte auf eine erzwungene Unterscheidung verschiedener Anforderungstypen verzichtet.

Werkzeuge im RE-Prozess

Ein klassischer Requirements-Engineering Prozess der PMT-Abteilung beginnt damit, das von einem OEM ein Pflichtenheft zu einem bestimmten Auftrag bei der HELLA eingeht. Dieses Lastenheft erhält die PMT-Abteilung üblicherweise in Form einer *DOORS*-Datei (Dynamic Object Oriented Requirements System).

DOORS ist ein Werkzeug zur Hilfe beim Requirements-Engineering und speichert Anforderungsdokumente vergleichbar zu einer herkömmlichen Datenbank. Dort wird in einer Tabellenstruktur der gesamte Inhalt eines Anforderungsdokumentes hinterlegt. Die einzelnen Anforderungen sind dabei zeilenweise angeordnet. Zusätzlich befinden sich in verschiedenen Spalten ergänzende Informationen zu der jeweiligen Anforderungen, wie z.B. die eindeutige Identifizierungsnummer oder die Unterscheidung zwischen Anforderung und genereller Information.

DOORS bietet zusätzlich eine Vielzahl von Plugins zur Verarbeitung, sowie die Möglichkeit, das DOORS-Objekt als PDF-Datei zu exportieren. Da man Informationsverlust verhindern möchte werden auch alle Zusatzinformationen aus den Spalten der Tabelle in der extrahierten Datei mit aufgenommen, was jedoch zu stark fragmentierten Dokumenten bzw. Texten führt. Dies ist unter anderem ein Problem, welches zu einem späteren Zeitpunkt noch einmal aufgegriffen wird.

Eines der oben erwähnten Plugins ist das von HELLA entwickelte *ReqPat*. Dieses wird dazu verwendet Teile von Anforderungsdokumenten auf Konsistenz zu überprüfen. Dazu werden sogenannte *Boilerplates* als Satz-

schablonen verwendet um die natürliche Sprache der Anforderungen in eine formalisierte Form zu bringen. Daraus kann dann ein prüfbarer Graph erstellt werden. Grundsätzlich bestehen hier parallelen zu dem später vorgestellten Konzept, jedoch ist ReqPat dabei auf bestimmte Anforderungen beschränkt.

Die Anforderungen aus dem DOORS-Objekt werden nun von den Mitarbeitern von PMT überprüft, ggf. an Firmen-interne Standards oder eine formalisierte Form überführt und dann an die verantwortliche Abteilung freigegeben. Dieser Ablauf ist in der Abbildung X.X präziser dargestellt.

HIER DIE ABBILDUNG INPUT-REQS -> FREIGEgeben

Dies kann etwa mit dem beschriebenen ReqPat geschehen. Da dieses Werkzeug jedoch nur einen Teil der bestehenden Anforderungen abdecken kann, muss der Rest der Lastenhefte oftmals manuell verarbeitet werden. Daraus entsteht enormer Aufwand und die Gefahr nicht explizit ausformulierte Abhängigkeiten zu übersehen und somit Fehler in den freigegebenen Anforderungen zu erhalten.

3.3 Anforderungen im RE-Management

Dieses Kapitel beschäftigt sich mit der Motivation und den Rahmenbedingungen für die später in dieser Arbeit skizzierte Werkzeugunterstützung. Wesentliche Erkenntnisse entstanden dabei in Rücksprache mit dem Requirements-Experten Colin Hood, der als externer Beschäftigter für die PMT-Abteilung von HELLA arbeitet.

Auf dieser Basis konnten Anforderungen an unser Tool für den Einsatz zur Transformation von Projektlastenheften während des RE erhoben werden.

Mithilfe unseres Werkzeugs kann, wie später gezeigt wird, das Anforderungsmanagement in Unternehmen maßgeblich verbessert werden. Das Ziel ist es, die im vorigen Abschnitt benannten Komplikationen beim Umgang mit Anforderungen zu bewältigen.

3.3.1 Kriterien der Qualität von Anforderungen

Das Ziel von Abteilungen, die sich mit Requirements-Engineering in Unternehmen beschäftigen, ist die Verbesserung der Qualität von Anforderungsdokumenten für die spätere Verwendung. Die Requirements-Experten sind

meist zwar nicht direkt am Entwicklungsprozess beteiligt, unterstützen die Experten in der Entwicklung jedoch beim Verständnis und der Umsetzung der Pflichtenhefte.

Wie im vorigen Abschnitt beschrieben, entsteht bei den Entwicklern auf Seite des Zulieferers häufig mangelndes Verständnis der Anforderungen des OEMs, wenn dort mit der Entwicklung begonnen werden soll. Diese Schwierigkeiten entstehen durch ein unterschiedliches bzw. mangelndes Verständnis der Anforderungen zwischen OEM und Zulieferer während des Entwicklungsprozesses. Die Anforderungen sind meist unvollständig spezifiziert, was zu Inkonsistenzen im Pflichtenheft führt. Requirements-Experten beim Zulieferer zielen also auf eine Analyse von Abhängigkeiten zwischen Anforderungen des Pflichtenheftes für die nachgelagerte Entwicklung ab.

Qualitätsmerkmale

In diesem Zusammenhang existieren Qualitätsmerkmale, die gemäß [ZG02] wesentliche Kriterien für die Bewertung von Anforderungen darstellen. Requirements-Experten orientieren sich dabei an der Unterscheidung der drei wesentlichen Qualitätsmerkmale *Consistency-Correctness-Completeness*, die sich wie folgt auffassen lassen:

1. *Consistency - Einheitlichkeit* Die Anforderungen dürfen sich an verschiedenen Stellen im Pflichtenheft nicht widersprechen, sondern sollen sich vielmehr an geeigneten Stellen ergänzen. Beispielsweise darf später keine Anforderung im Pflichtenheft enthalten sein, die im direkten Widerspruch zu anderen Anforderungen steht, die bereits zuvor anders spezifiziert wurden. Entwickler, die nur die für sie relevanten Stellen des Pflichtenheftes erhalten, würden somit fundamental unterschiedliche Produkte entwerfen. Inkonsistenzen und Abhängigkeiten zwischen Anforderungen bedürfen also Rücksprache mit dem OEM bei der Bewertung.
2. *Completeness - Vollständigkeit* Es sollten zu allen relevanten Merkmalen des Endproduktes entsprechende Anforderungen existieren. Das Pflichtenheft sollte also alle Produktmerkmale inhaltlich abdecken, die für die Entwicklung beim Zulieferer relevant sind. Wenn der OEM etwa konkrete Vorstellungen zum Design des Produktes hat, diese jedoch im Pflichtenheft fehlen, erhält der Zulieferer diese Informationen nicht (etwa implizite Informationen oder Erläuterungen). Es gehen mitunter

entscheidende Informationen und Kriterien für den Entwicklungsprozess verloren.

3. *Correctness - Richtigkeit* [ZG02] beschreibt die Richtigkeit von Anforderungen als ein prinzipiell vages Konzept aus zwei verschiedenen Standpunkten:
 - (a) Formal betrachtet ist ein Anforderungsdokument korrekt, wenn es einheitlich und vollständig ist, d.h. frei von Widersprüchen und inhaltlich alles beinhaltet, was als "wahr" über das Produkt bezeichnet werden kann.
 - (b) Praktisch erfolgt die Bewertung der Richtigkeit eines Dokumentes anhand von Kriterien, die Stakeholder auswählen. Das Pflichtenheft wird demnach als richtig empfunden, wenn es Bedürfnisse zufriedenstellend abbildet. Aus Sicht der Kostenkalkulation wird etwa beurteilt, ob die Kosten durch die zu verwendenden Materialien, die im Pflichtenheft vermerkt sind, gering gehalten werden.

Es lassen sich gemäß [ZG02] weitere Abhängigkeiten zwischen den drei Merkmalen beobachten, die die Umsetzung in konkreten Dokumenten erschweren können. 1. und 2. stehen demnach häufig in Wechselwirkung zueinander.

Zur Herstellung von Konsistenz im Pflichtenheft existiert der Ansatz, einzelne Anforderungen entweder schwerer zu gewichten als andere, oder unpassende Anforderungen komplett zu entfernen. Dies führt jedoch zu verminderter Vollständigkeit.

Informationsmangel kann zwar ebenfalls recht einfach durch das Hinzufügen neuer Anforderungen ausgeglichen werden, jedoch muss dabei darauf geachtet werden, dass die Spezifikation dann nicht zu umfänglich ausfällt. Verbesserung der Vollständigkeit kann somit auch die Richtigkeit des beschriebenen Produktes beeinflussen, sollten entsprechende Informationen ergänzt werden.

Beispiel problematischer Anforderungen

Zur Verdeutlichung des resultierenden Aufwands bei der Herstellung hochqualitativer Anforderungen soll das folgende fiktive Beispiel dienen. Die Analyse der Vollständigkeit soll hier vernachlässigt werden, da es sich nur um einen fiktiven Ausschnitt handelt. Folgende funktionale Anforderungen könn-

3.3. ANFORDERUNGEN IM RE-MANAGEMENT Analysis von Projektlastenheften

ten in einem Pflichtenheft enthalten sein:

- 1: Design: Das Gehäuse soll rot gestrichen sein.
- ...
- 2: Materialien: Das Gehäuse soll nach außen aus 1mm starkem, rostfreiem Stahl bestehen.

Bei der Prüfung auf Vollständigkeit fallen zunächst keine Konflikte auf, jedoch ist 1recht vage formuliert. Genaue Angaben zur verwendeten Farbe und der genaue Farbton fehlen in der Anforderung. Dieses Problem wird bei der Analyse der Abhängigkeiten offensichtlich, da die Anforderung rostfreier Stahl"die Verwendung von entsprechender Farbe notwendig macht.

Eine solche Anforderung könnte etwa wie folgt lauten:

- ...
- 3. Materialien: Die Lackierung erfolgt mit Metallschutzlack.

Isoliert betrachtet wäre diese Anforderung auch für alle Lackierungen denkbar, nicht nur für das Gehäuse. Daher muss das Ausmaß dieser Anforderung auf das Gehäuse beschränkt werden, etwa durch 1.1 als zugehörig zu 1.

Aus Perspektive der Richtigkeit könnten je nach Stakeholder, der die Bewertung vornimmt, weiter spezifiziert werden. Ein Designer könnte etwa einen genauen Farbcode zur Farbanforderung „rot“ hinzufügen. Aus Perspektive der Kosten könnte 2auch komplett gegenüber einer günstiger umzusetzenden Anforderung entfallen, die etwa „Kunststoff“ anstelle von „rostfreiem Stahl“ vorschreibt.

Problematisch ist bei der Betrachtung des Beispiels auch, dass in der Regel interdisziplinäre Teams jeweils nur einen Abschnitt des Pflichtenheft erhalten und ausdesignen. Sie verfügen somit nur über einen Teilbereich des Wissens bzw. nur einen Teil der Anforderungen. Es könnte also der Fall eintreten, dass das Design-Team nur 1 und das Ingenieur-Team nur 2 erhält.²

Die Relevanz dieser drei Kriterien für das betriebliche Anforderungsmanagement wurde auch in Rücksprache mit internationalen Requirements-Experten bei HELLA betont. Auf dieser Basis realisieren die dortigen Teams

²Hier ist kurz ein naiver Ansatz zur Verdeutlichung der Komplikationen beschrieben. [ZG02] beschreibt detailliert die Prüfung der Qualitätsmerkmale anhand verschiedener Verfahren und wie diese verbessert werden können. Mithilfe von Werkzeugen können diese auf Basis automatisierter Textanalyse der Anforderungen umgesetzt werden.

Analysen der Projektlastenhefte.

3.3.2 Kriterien zur Herstellung qualitativer Anforderungen

Harant und Poethen stellen in [HP12] weitere Maßnahmen für das Requirements-Management auf Basis des *HOOD Capability Models* (HCM) vor. Dabei ordnen sie die verschiedenen Schritte anhand eines mehrstufigen Prozesses ein, der in jedem Entwicklungsprojekt von Requirements-Experten durchgeführt werden soll (siehe Abb. 3.1).

Diese Expertengruppe dient dabei abstrakt zur Unterstützung der tatsächlichen Entwicklungsarbeit. Weitere Merkmale von qualitativen Anforderungen werden durch das HCM beschrieben und verschieden gewichtet.



Abbildung 3.1: Überblick der 3 Qualitätsstufen von Anforderungen im *Requirements-Management & Engineering* (RM & E) gemäß [HP12]

Dabei unterscheidet das Modell zwischen 3 verschiedenen Qualitätsstufen von Anforderungen. Diese werden jeweils durch verschiedene Kriterien dargestellt. Das Ziel der Arbeit der Abteilung bei HELLA stellt demnach das Herstellen einer möglichst hohen Qualitätsstufe der Pflichtenhefte dar.

Im besten Fall erreichen die Experten dort die Stufe *Expert* und machen die Anforderungen somit *vollständig & verfolgbar*.

3.3.3 Umgang mit natürlicher Sprache in Lastenheften

Unvollständige Anforderungen lassen sich meist auf falsche Annahmen der Stakeholder beim OEM zurückführen, die implizite Annahmen bei der Formulierung von Anforderungen treffen und diese Annahmen nicht explizit erneut aufführen. Ergänzenden Informationen zu Anforderungen können jedoch, wie zuvor gezeigt, durchaus andersartig kategorisiert im Pflichtenheft vermerkt sein.

Dies resultiert laut [PR15] daher, dass die verschiedenen Autoren des Pflichtenheftes beim OEM als selbstverständlich erachten und diese nicht nochmal für externe Leser hinzufügen. Es stellt sich jedoch die Frage, ob unterschiedliche Menschen jemals ein verlässlich gleiches Verständnis der gleichen Sache haben.

Colin Hood formuliert Anforderungen grundsätzlich als „individuelle Bedürfnisse“, die immer abhängig von dem Autoren sind.³ In wird die Verwendung von Sprachschablonen für Projektlastenhefte im Vergleich zu natürlicher Sprache evaluiert. Die grundsätzliche Eignung natürlicher Sprache für die Formulierung von Anforderungen wird dort bezweifelt.

Hood stellte die Eingrenzung der Sprache, die für Projektlastenhefte genutzt werden darf, als eine vielversprechende Möglichkeit vor, sprachliche Eigenheiten der Autoren aus verschiedenen Bereichen effektiv zu vermeiden. Diese Transformation natürlichsprachlicher Anforderungen in formale Sprache als Schablonen stellt auch das Ziel unserer Arbeit dar.

3.4 Anwendungsfall bei HELLA

Der Anwendungsfall für die zu entwickelnden Werkzeuge ist ähnlich zu dem oben beschriebenen. Es wird davon ausgegangen, dass die HELLA einen Auftrag für ein Produkt von einem OEM erhält. Es ist jedoch bekannt, dass ein ähnliches Produkt, etwa eine frühere Version, bereits von HELLA entwickelt wurde.

³Im Rahmen unserer Arbeit konnten wir mehrmals Interviews zur Ausrichtung bzw. dem Ansatz unserer Prototypen führen.

Somit liegt ein Pflichtenheft eines OEM in natürlicher Sprache vor. Außerdem existiert zu dem alten Produkt der HELLA bereits in Lastenheft in formalisierter Form.

Ziel ist es die beiden Anforderungsdokumente zu vergleichen um alle Unterschiede der beiden Produkte zu ermitteln. Diese könnten dann etwa genutzt werden um eine sehr frühe und recht präzise Kostenabschätzung der neuen Produktentwicklung an den OEM senden zu können.

Hierzu muss jedoch zunächst das Pflichtenheft des OEM in eine vergleichbare formalisierte Form gebracht werden. Erst sobald sich also beide Anforderungsdokumente in einer gleichartigen Form befinden kann ein direkter Vergleich der von den Dokumenten beschriebenen Produkte stattfinden.

Diese Heterogenität von Anforderungsdokumenten ist in der Automobilindustrie vor allem bei der Beziehung zwischen OEM und Zulieferer ein bekanntes und wiederkehrendes Problem.

Daher wird zunächst ein Werkzeug benötigt, welches mit kleinstmöglichem Nutzeraufwand und größtmöglicher Sicherheit ein natürlichsprachliches Anforderungsdokument formalisieren kann.

Verarbeitet man mit diesem Werkzeug das Pflichtenheft des OEM, so erhält man zwei homogene Dokumente. Hierzu soll außerdem ein Werkzeug entwickelt werden, welches in der Lage ist zwei solcher Dokument zu vergleichen und Unterschiede darzustellen.

HIER USE-CASE ABBILDUNG VON FIRMEN POWERPOINT

Die Ansätze und Konzepte solcher Werkzeuge sind in dem folgenden Abschnitt genau beschrieben.

3.5 Ansatz und Konzept der Werkzeuge

Wie bereits unter beschrieben, existiert eine Vielzahl an Arbeiten, die sich mit der verlässlichen Überführung der Anforderungen in natürlicher Sprache in Sprachschablonen beschäftigen. Dazu werden Methoden aus dem NLP und Ontologien genutzt, um den Text der Anforderungen zunächst genau zu annotieren. Dieser Ansatz stellt die Basis für die Vorstellung unseres Werkzeugs in diesem Kapitel dar.

In [ZH17] wurde das Potential einer Verbesserung von RE-Prozessen in

Firmen durch Werkzeuge aufgezeigt. Die in den vorigen Kapiteln aufgezeigten Anforderungen an das Requirements-Management schaffen somit die Basis für den konkreten Entwurf von Werkzeugen, die entsprechende Funktionalitäten bereitstellen.

3.5.1 Anforderungen an die Entwicklung

Für den Entwurf unserer Prototypen stellt sich einerseits die Frage, an welcher Stelle im abstrakten Modell ein Potential für Werkzeugunterstützung entsteht, auf der anderen Seite aber auch die Definition einer konkreten Zielgruppe bzw. Stakeholdern im tatsächlichen Betrieb bei HELLA, die nachher mit der Software interagieren und diese täglich nutzen. Für die Entwicklung unseres Werkzeugs haben wir in Zusammenarbeit mit RE-Experten einige Ziele identifiziert, die eine Unterstützung für die Arbeit ausmachen würden:

1. Zeitersparnis bei Arbeit mit dem Werkzeug, da der Aufwand einer manuellen Prüfung reduziert wird
2. Hoher Automatisierungsgrad bei der Verarbeitung, um Konflikte in den Anforderungen im Voraus aufzulösen
3. Hoher Abdeckungsgrad verschiedener Problemfälle, da eine einfache Wiederverwendbarkeit den Aufwand zur Anpassung an andere Domänen reduziert
4. Mehrwert für den Nutzer, da effektiv Arbeitszeit gespart und Aufwand an das Werkzeug abgegeben werden kann
5. Anforderungen an NLP-Werkzeuge umsetzen, da eine präzise sprachliche Analyse eine wesentliche Grundlage für die Verlässlichkeit der Verarbeitung darstellt.

3.5.2 Zuordnung einzelner Werkzeuge

Verschiedene Verfahren, bereits bei der Spezifikation von Anforderungen im Zusammenhang eines Pflichtengeftes durch Werkzeuge zu unterstützen, sind unter VERWANDTE AREITEN beschrieben. Es existieren jedoch durch die Besonderheiten im geteilten Entwicklungsprozess zwischen OEM und Zulieferer besondere Herausforderungen für das RE, welche Potential für eine Werkzeugunterstützung bieten.

Wie bereits in Kap. X.X beschrieben, resultieren die Probleme auf Seiten des Zulieferers beim Verständnis der Anforderungen aus dem Umstand, dass dort ein Pflichtenheft in natürlicher Sprache vorliegt. Dieses verursacht einerseits einen Arbeitsaufwand durch die Probleme beim Umgang mit natürlicher Sprache wie z.B. sprachlichen Mehrdeutigkeiten (Kap. X.X). Andererseits können enthaltenen Anforderungen teilweise nicht korrekt bzw. von niedriger Qualität sein, was zu inhaltlichen Problemen für die Produktentwicklung führt.

Auch kann es mehrere Versionen der Spezifikation eines Produktes im Laufe der Zeit geben, deren Unterschiede erkannt werden müssen. Auswirkungen der Überarbeitung des Produktes auf Basis einer neuen Version des Pflichtenheftes müssen unter verschiedenen Kriterien gewichtet werden.

Es gilt also, Werkzeuge zu entwerfen, die insbesondere bei der Lösung von sprachlichen Problemen und bei der Identifikation von Konflikten zwischen Anforderungen in Pflichtenheften unterstützen.

Daher werden im folgenden die Ansätze aus [ZH17] beschrieben, die eine Transformation von bereits spezifizierten Anforderungen in bestehenden Pflichtenheften in eine besser verständliche und eindeutige formale Sprache vorschlagen. Dieser Schritt soll für jedes neu eintreffende Pflichtenheft durchgeführt werden, damit dieses dann standardisiert vorliegt. Die zweite Software-Komponente dient zum automatisierten Vergleich mehrerer verschiedener Pflichtenhefte unter der Verwendung von Ontologien.

Diese Ansätze haben zur konkreten Entwicklung von zwei Werkzeugen geführt, dem *Requirements-To-Boilerplate-Converter* (R2BC) und dem *Delta-Analyser* (DA). Diese identifizierten Software-Artefakte werden für die Verwendung in einer prototypischen Werkzeugkette hintereinander angeordnet. Somit erfolgt auf Basis der Konvertierung in Schritt 1 die Delta-Analyse in Schritt 2.

R2BC

Zu Beginn des Verarbeitungsprozesses liegen die Pflichtenhefte in natürlicher Sprache vor, die der Zulieferer vom Kunden erhalten hat. Im ersten Schritt werden diese natürlichsprachlich formulierten Anforderungen durch den R2BC in Sprachschablonen überführt.

Nach der Verarbeitung liegt das Pflichtenheft des OEM somit in formaler Sprache vor. Diese Version des Pflichtenheftes kann dann für den Entwicklungsprozess genutzt werden. Das Ziel des R2BC ist es also, die Lesbarkeit

bzw. die Eindeutigkeit der Anforderungen durch Formalisierung zu verbessern. Auf dieser Basis kann dann der Inhalt des Pflichtenheftes eindeutig durch die Entwickler beim Zulieferer verstanden werden. Der Prozess ist in Abb. 3.2 dargestellt.

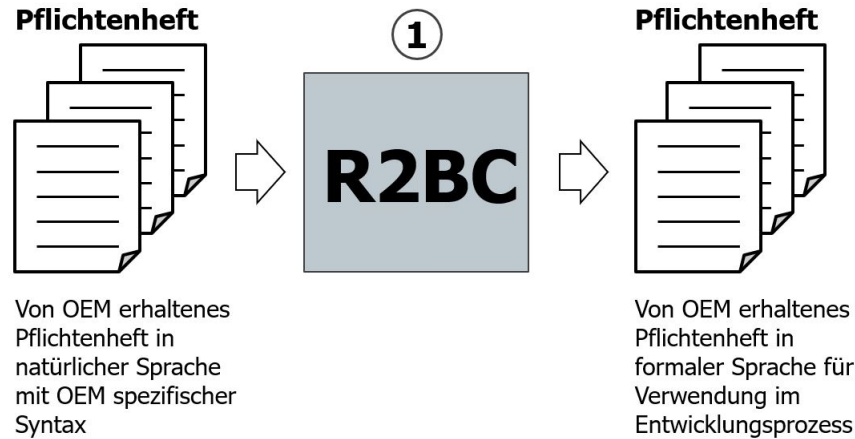


Abbildung 3.2: Szenario für Unterstützung durch R2BC-Komponente [eigene Darstellung]

DA

Diese logisch eindeutige Form der Anforderungen ermöglicht auch einen Vergleich mehrerer Pflichtenhefte anhand der Struktur der einzelnen Anforderungen in einer Ontologie. Diese Unterschiede können durch den DA identifiziert und in Form eines Deta-Reports ausgegeben werden. Unterschiede können anschließend etwa für einen RE-Experten auf dem Bildschirm bzw. in eine Datei ausgegeben werden. Der Ablauf ist in Abb. 3.3 dargestellt.

Alternativ können Änderungen auch automatisiert durch eine Integration der Anpassungen in das interne Pflichtenheft übernommen werden. Dazu müssen im Programm Verfahren zur Gewichtung der Deltas bzw. Metriken zur Entscheidung zwischen Anforderungen hinterlegt sein.

Angesichts der Aufgaben für die Werkzeuge stellt sich die Frage nach dem erreichbaren Grad der Automatisierung bei hoher Verlässlichkeit im Gegensatz zu möglichen erforderlichen Interaktionspunkten mit dem Nutzer.

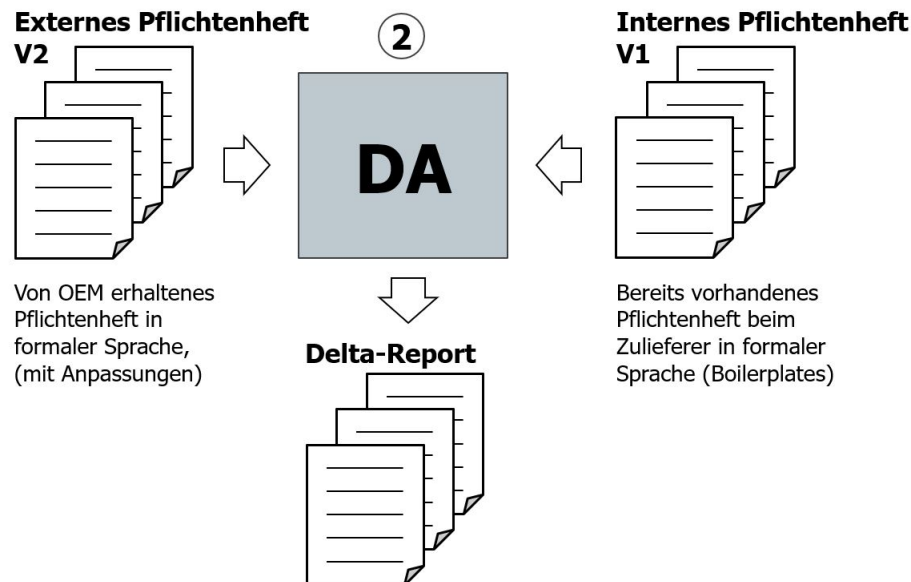


Abbildung 3.3: Szenario für Unterstützung durch DA-Komponente [eigene Darstellung]

Der prinzipielle Aufbau beider Werkzeuge wird in den folgenden Abschnitten beschrieben.

3.5.3 R2BC-Funktionalität

Wie bereits zuvor beschrieben wurde, stellt der *Requirements-To-Boilerplate-Converter* die erste wesentliche Stufe bei der automatisierten Verarbeitung von Pflichtenheften dar. Die Textverarbeitung erfolgt dabei sukzessive bzw. satzweise aus dem Quelldokument in ein Zieldokument, dass am Ende alle Anforderungen in angepasster Form enthält.

Dabei greift er auf eine formale Sprache zurück, die in verschiedenen Schablonen repräsentiert und im Programm gespeichert ist. Diese eingeschränkte Sprache ist Zulieferer-intern fest dokumentiert und wurde zuvor durch RE-Experten eindeutig definiert, was zu einer verbesserten Verständlichkeit führt. Sprachschablonen stellen demnach, im Kontrast zu natürlicher Sprache, eine eindeutige Deutung der Anforderungen sicher. Im besten Fall wurden somit insbesondere sprachliche Mehrdeutigkeiten im Sinne der Ato-

marität von Anforderungen eliminiert bzw. aufgeschlüsselt.

Der R2BC verarbeitet die Pflichtenhefte unter Zugriff auf diese formalen Sprachschablonen (*Boilerplates*) automatisiert und ein Eingreifen des Nutzers ist nur im Ausnahmefall erforderlich.

Programmablauf und Artefakte

Auf Basis dieses Ansatzes wurde ein Algorithmus zur satzweisen Verarbeitung der Sätze unter Verwendung des NLP-Werkzeugs GATE entwickelt, der aus natürlichsprachlichen Eingaben (Input) eine Ausgabe (Output) in formaler Sprache erzeugt. Dabei werden auch Metainformationen über den Text generiert und zur weiteren Verarbeitung ausgegeben.

Der prinzipielle Programmablauf gliedert sich wie folgt:

1. Zunächst werden die Sätze für eine sukzessive Verarbeitung getrennt. Dies erfolgt mithilfe des NLP-Werkzeugs GATE.
2. Für jeden Satz wird geprüft, ob und in welchem Maße Anpassungen an die formale Sprache der Boilerplates erforderlich sind. Dazu werden Syntax bzw. Satzbestandteile analysiert.
3. Passende Boilerplate-Typen für die Anforderung wird identifiziert. Dieser leitet sich aus den unter 2. gefundenen Merkmalen ab. Es können etwa funktionale von nicht-funktionalen unterschieden sowie Bedingungen für die Anforderung modelliert werden.
4. Sind Anpassungen erforderlich, so wird eine Konvertierung in die jeweils passende Boilerplate durchgeführt und diese werden im Programm gesammelt.
5. Die erzeugte Boilerplate, die am besten zum Inhalt der Anforderung passt, wird identifiziert. Dies kann automatisiert oder unter Einbezug des Nutzers geschehen. Der Nutzer kann dazu beispielsweise aus mehreren Vorschlägen auswählen.
6. Sind alle Anforderungen aus dem Quelldokument verarbeitet, wird auf Basis der erzeugten Anforderungen in Boilerplates ein Zieldokument für die weitere Verwendung erzeugt.

Auf Basis dieses Konzeptes wurde in QUELLE PAPER ein Vorschlag zur Aufteilung der Ressourcen in verschiedene Bereiche gemacht. Dabei gliedern sich die Funktionalitäten in drei Software-Komponenten, die in Abb. 3.4 dargestellt sind. Der Programmablauf ist durch die Schritte 1 und 2 grob

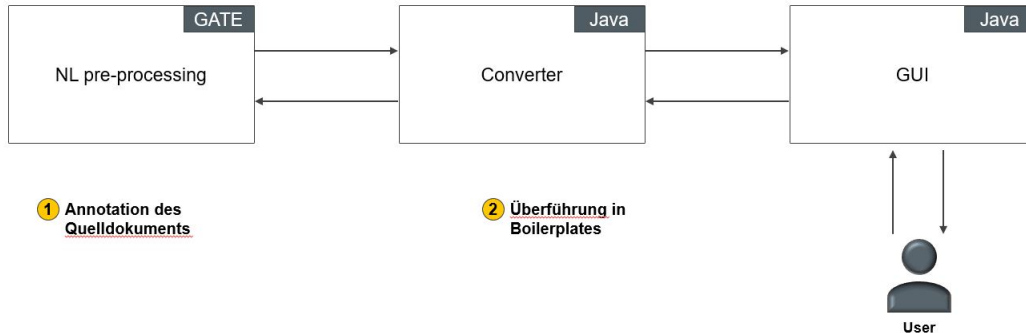


Abbildung 3.4: Verteilung der R2BC-Ressourcen [eigene Darstellung]

dargestellt und verteilt sich hauptsächlich auf die zwei entsprechenden Artefakte. Die Komponenten lassen sich dabei wie folgt definieren:

- Die *NL pre-processing* Komponente stellt das backend für die Sprachverarbeitung des Quelldokumentes dar. Dabei werden verschiedene Textbereiche mithilfe des NLP-Werkzeugs GATE annotiert und somit relevante Informationen über den Text generiert. Das Ziel ist es, konkrete Anforderungen im Pflichtenheft eindeutig zu identifizieren. GATE stellt ein in hohem Maße konfigurierbares Werkzeug auf *java*-Basis dar, dass in Kap. 4.1 näher beschrieben wird.
- Die zweite Komponente beinhaltet die Konvertierung der Sätze aus dem Quelldokument in Sprachscablonen. Der *Converter* nutzt dabei die zuvor erstellten Annotationen, um für jeden Satz im Anforderungsdokument eine oder mehrere passende Sprachschablonen zu finden. Diese Schablonen werden dann durch das Programm generiert bzw. passend „befüllt“. Der *Converter* wird ebenfalls mit *java* implementiert.
- Die Interaktion mit dem Nutzer wird durch ein herkömmliches *GUI* realisiert, dass dem Nutzer die Navigation im Programm ermöglicht. Auch kann dort etwa die Anzeige und Auswahl von einer aus mehreren möglichen Konvertierungen erfolgen.

In Kap. 4 erfolgt auf dieser Basis die konkrete Implementierung eines Prototypen für den R2BC.

3.5.4 DA-Funktionalität

Auf Basis des zuvor beschriebenen R2BC stellt der *Delta-Analyser* die Softwarekomponente für einen kriteriengeleiteten Vergleich zweier Anforderungsdokumente dar. Das Ziel ist es, auf Basis der Informationen zu jeweils einem der Lastenhefte den Grad der inhaltlichen Überschneidung zum anderen Lastenheft zu messen.

Im Kontext des Use-Cases dieser Arbeit geht es dabei insbesondere um den Vergleich von zwei Versionen des gleichen Anforderungsdokumentes, wenn auf Unterschiede in der Überarbeitung geprüft werden soll. Zu Anforderungen aus dem Lastenheft V2 müssen also inhaltlich passende Anforderungen aus Lastenheft V1 gefunden werden, um anschließend Deltas zwischen diesen Anforderungen zu identifizieren bzw. für den Nutzer hervorzuheben. Diese Verarbeitung erfolgt dabei satzweise für das komplette Lastenheft.

Für ein automatisiertes Finden passender Requirements in unserem Werkzeug müssen daher die Anforderungen in einheitlicher Form dargestellt sein, damit verlässlich nach einzelnen Satz- bzw. Textbestandteilen in den Anforderungen gesucht und verglichen werden kann. Dabei stellt, wie zuvor beschrieben, die Überführung der Anforderungen aus den Dokumenten aus der natürlichen in eine fest definierte, formale Sprache, eine wesentliche Grundlage dar. Der zuvor skizzierte R2BC bietet eben diese Funktionalität.

Programmablauf und Artefakte

Die beschriebenen Funktionen ermöglichen den Entwurf eines Verfahrens für einen automatisierten Vergleich von zwei Lastenheften im DA. Dazu wird jede Anforderungen aus dem Lastenheft V2 mit denen aus V1 abgeglichen, um jeweils zueinander passende zu finden. Folgendes Vorgehen wurde dazu entwickelt:

1. Die einzelnen Anforderungen in Boilerplates werden für Lastenheft V1 und V2 geladen.
2. Für jede Boilerplate aus Lastenheft V2 wird geprüft, ob es identische, vergleichbare oder keine zuordbaren Anforderungen in Lastenheft V1

gibt. Diese drei Kategorien werden auf Basis der GATE-Annotationen aus dem R2BC identifiziert.

3. Entsprechend des gefundenen Übereinstimmungsgrades werden die Deltas zwischen den Anforderungen identifiziert:

- (a) *Identische Anforderung*: Es wurde eine identische Anforderung in beiden Lastenheften gefunden. Die Anforderung kann ohne gesonderte Betrachtung übernommen werden.
- (b) *Vergleichbare Anforderung*: Es wurden nur eine teilweise übereinstimmende Anforderung gefunden. Dies ist etwa der Fall, wenn sich Anforderungen geändert haben bzw. überarbeitet wurden. Zwischen diesen Anforderungen existiert also ein Delta, das eine Änderung in den Anforderungen signalisiert und somit den Produktspezifikationen bedeuten würde.

Eine feingranulare Aufschlüsselung der erwartbaren Deltas findet sich in der Implementierung in Kap. 4.x. Dort werden auch Vorschläge zur Gewichtung verschiedener Deltas im Kontext des RE gemacht.

- (c) *Keine vergleichbare Anforderung*: Es konnte keine Anforderung aus V1 gefunden werden. In Lastenheft V2 könnten etwa neue Anforderungen enthalten sein, die in V1 noch in keiner Weise spezifiziert wurden. Dies kann etwa Eigenschaften betreffen, die erst später im Entwicklungsprozess definiert wurden oder zu der vorher noch keine Informationen bekannt waren. Im Kontext anderer Anforderungen muss dabei auf Konsistenz geprüft werden.
4. Falls es noch mehr mindestens vergleichbare Anforderungen in V1 gibt, wird die am besten passende Anforderung gefunden.
 5. Der Delta-Report für den Nutzer wird unter Zugriff auf die Ergebnisse aus 4. erzeugt, wenn V2 vollständig durchlaufen wurde.

Insofern Deltas gefunden wurden, muss immer auch die Umsetzbarkeit dieser neuen Anforderungen gesondert geprüft werden. Die weitere Korrektheit des Lastenheftes V2 zu V1 kann bei vielen Deltas nicht gewährleistet werden, was eine Umsetzung der Anforderungen erschwert. Es können massive Unterschiede in den Spezifikationen auftreten, was zu einer teilweisen oder kompletten Neuentwicklung des Produktes führen kann.

Die beschriebene Funktionalität hat in der Entwicklung unseres Werkzeugs zur konkreten Herausbildung von drei Kernbestandteilen geführt. Unser Werkzeug beinhaltet dabei Komponenten zur Abbildung von Fachlogik und GUI. Die Verteilung der Ressourcen ist in Abb. 3.4 dargestellt. Die Kom-

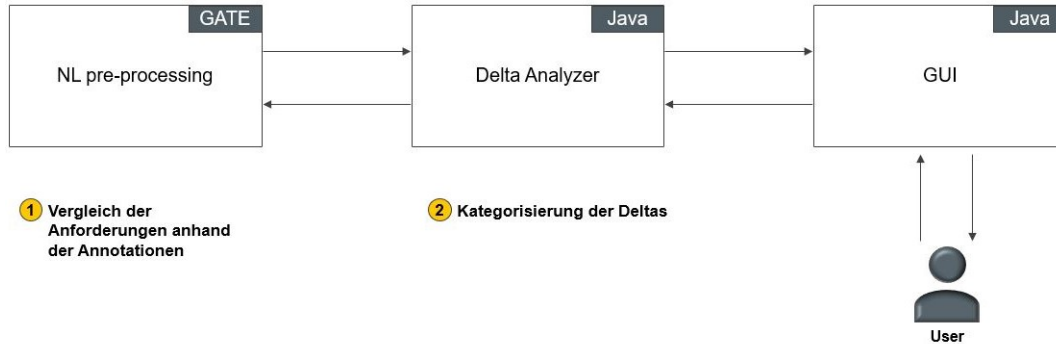


Abbildung 3.5: Verteilung der DA-Ressourcen [eigene Darstellung]

ponenten realisieren dabei die zwei Hauptfunktionalitäten 1 und 2 in Form von GATE und java-Ressourcen. Auf dieser Basis können den einzelnen Komponenten Funktionalitäten wie folgt zugeordnet werden:

- *NL-pre-processing* Wie schon beim R2BC dient diese Komponente zur Generierung von Informationen über den annotierten Text. Da der DA die zweite Stufe der Verarbeitung darstellt, nutzen wir hierzu die Ergebnisse aus dem R2BC, d.h. die dort generierten Annotationen, die wir in den DA einlesen. Somit können die Anforderungen, die nach dem R2BC in Boilerpaltes vorliegen, anhand ihrer Annotaionen verglichen werden.
- Der *Delta-Analyser* prüft die verschiedenen Anforderungen aus Pflichtenheft V2 auf ihren Übereinstimmungsgrad mit den Anforderungen aus V1. Dabei greift er auf die geladenen Annotationen beider Pflichtenhefte zu und vergleicht daran verschiedene Anforderungen. Schlussendlich wird ein Delta-Report erstellt, der alle Deltas zwischen den Anforderungen enthält.
- Die *GUI* ermöglicht dem Nutzer die Eingabe der zwei zu vergleichenden Pflichtenhefte. Nach der Verarbeitung erhält der Nutzer dort eine

Übersicht der Unterschiede der beiden Pflichtenhefte, indem die jeweils verglichenen Anforderungen dort angezeigt werden.

Die Implementierung des DA auf Basis des zuvor implementierten R2BC ist in Kap. 5 beschrieben. Dort werden auch konkrete Erweiterungen und Verbesserungspotentiale des Programms beschrieben, die nach der Entwicklung identifiziert wurden.

Kapitel 4

Requirements-To-Boilerplate-Converter (R2BC)

Dieses Kapitel behandelt den *Requirements-to-Boilerplate-Converter* (*R2BC*) ausführlich. Hierzu wird zunächst der Aufbau des Tools aus den verschiedenen Klassen und der verwendeten Software beschrieben. Daraufhin wird die Implementierung wichtiger Funktionalitäten und Datenstrukturen mit Auszügen des Quellcodes gezeigt und erläutert. Anschließend wird das Testverfahren für das Werkzeug beschrieben und die dadurch erzielten Ergebnisse aufgeführt. Die Evaluation der Ergebnisse findet in einem späteren Kapitel statt. Abschließend wird ein Katalog mit möglichen Erweiterung und Verbesserungen für das Werkzeug geliefert, welche auf Grund mangelnder Zeit und Ressourcen von uns nicht mehr umgesetzt werden konnten.

4.1 Architektur, Klassen und Verteilung der Ressourcen

In diesem Abschnitt soll die Architektur des R2BC vorgestellt und erläutert werden. Dazu gehört die zur Entwicklung verwendete Software, sowie ein Überblick über die Klassenstruktur des Werkzeugs.

4.1.1 Verwendete Software

Als verwendete Software werden sämtliche Programme aufgeführt, welche eine Hilfe bei der Implementierung oder Teile der Funktionalität des Programmes darstellen.

Implementierungsunterstützung

Das Werkzeug wurde auf Basis der Programmiersprache JAVA geschrieben. Hierzu wurde das JDK (*JAVA development kit*) 8.1 verwendet.

Die Wahl der Programmiersprache fiel aufgrund von zwei Hauptfaktoren auf JAVA:

1. JAVA ist eine weit verbreitete und entwicklerfreundliche Programmiersprache. Vorteile der Sicherheit von JAVA werden außerdem in [RS18] aufgeführt.
2. Das Programm, welches für das NLP-Preprocessing verwendet wird ist JAVA-basiert. Somit fällt die Integration des Programms in unser Werkzeug einfacher. Dies wird jedoch noch genauer in einem späteren Abschnitt behandelt.

Als Programmierumgebung¹ wurde die Netbeans 8.2 IDE verwendet. Diese ist mit *Eclipse* eine der am weitesten verbreitete Programmierumgebung für JAVA-Programme.

Außerdem wurde die Software *Apache Maven* als Build-Management-Tool (*BMT*) verwendet. Dieses wird vor allem für JAVA-Programme verwendet und war daher sehr geeignet für die Verwendung in unserem Werkzeug. Außerdem wurde das selbe BMT auch in dem später beschriebene Tool zur Bereitstellung bestimmter Funktionalitäten verwendet. Build-Management-Tools helfen grundsätzlich dabei bestimmte, zuvor nicht vorhandene Bibliotheken und deren Funktionalitäten nachzuladen und in ein Programm zu integrieren. Außerdem helfen sie bei dem späteren Ausliefern des Programms, also der Erstellung eines unabhängig ausführbaren Programms außerhalb der Programmierumgebung. In unserem Fall wurde also eine *Executable JAR* (ausführbares JAVA-Archiv) erstellt.

¹oft als IDE, also *integrated development enviroment* bezeichnet

Funktionalitätsunterstützung

Das zuvor bereits erwähnte Programm zur Bereitstellung von Funktionen ist bekannt unter dem Namen GATE (*General Architecture for Text Engineering*). Dies ist ein Werkzeug zur Erstellung von NLP-Programmen und wird ausführlich in [RS18] behandelt.

In unserem Programm wird *GATE Embedded* verwendet. Dies ist eine Ansammlung von JAVA-Bibliotheken, welche alle Funktionen von GATE bieten. Man erhält also das gesamte Programm, nur ohne Benutzeroberfläche. Im folgenden wird daher im Zusammenhang auf die entwickelten Werkzeug nicht zwischen GATE und GATE Embedded unterschieden.

GATE ist, wie in Abbildung X.X (Konzept) zu sehen ist, für das NLP, also die Sprachverarbeitung in unserem Werkzeug zuständig und bildet somit das *Backend* des Programms.

Die genaue Umsetzung vom NLP in unserem Programm und der damit verbundene Einsatz von GATE sind jedoch Teil der Implementierung und werden somit in Abschnitt X.X behandelt.

4.2 Klassenstruktur

Im Rahmen der Entwicklung unseres Proof-of-concepts wurde die Funktionalität aus Kap. 3.x auf verschiedene Komponenten aufgeteilt. Hier soll lediglich ein grober Überblick der Ressourcenverteilung im Programm gegeben werden, der Programmablauf findet sich in Kap. 4.x.

Verschiedene Programmbausteine sind dabei modular aufgebaut und verteilen sich auf die Pakete *R2BC*, der die Fachlogik enthält, *R2BCGUI*, das eine simple GUI-Klasse enthält, und *Boilerplates*, das die Objekte zur Erzeugung von Sprachschablonen enthält (siehe Abb. 4.x).

Die Funktionalität wurde dabei hauptsächlich mit verschiedenen java-Klassen umgesetzt, die die Umsetzung der Fachlogik darstellen. Die Anbindung an das NLP-Werkzeug GATE-Embedded erfolgte dabei unter Verwendung verschiedener java-Bibliotheken, die extern im Programmordner gespeichert sind. Die GATE-Ressourcen wurden dabei zur Annotation von Text genutzt, die selbst entwickelten java-Klassen einerseits für die GATE-Konfiguration, andererseits für die Auswertung der dort erstellten Annotationen.

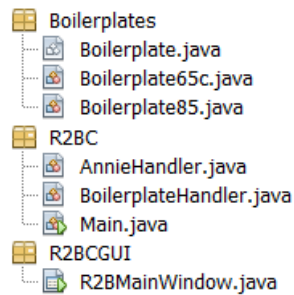
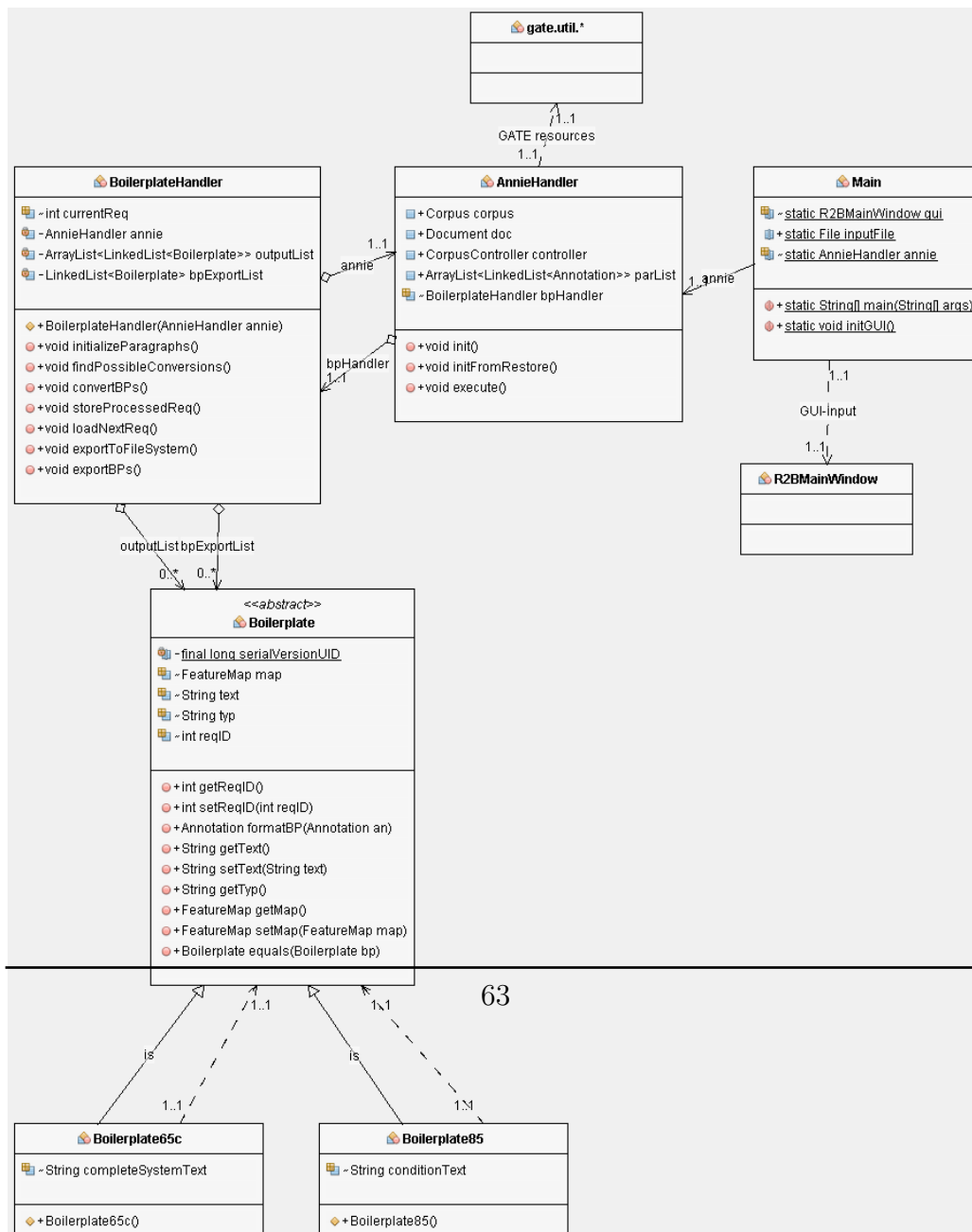


Abbildung 4.1: Zuordnung der R2BC-Klassen zu Paketen nach Abb.3.x [eigene Darstellung]



Auf Basis des in Kap. 3.x beschriebenen Algorithmus können die Aufgaben der verschiedenen Klassen in Abb. 4.x wie folgt erläutert werden:²

- *Main* - Die Hauptklasse ermöglicht den Zugriff auf alle Funktionen des Programms. Es erfolgt die Initialisierung aller Programmkomponenten. Es werden dazu verschiedene Objekte für die Textverarbeitung (*AnnieHandler*, *BoilerplateHandler*) sowie für die GUI (*R2BMainWindow*) erzeugt. Hauptsächlich dient die Hauptklasse auch der Initialisierung der GUI, um die dortigen Schaltflächen mit Funktionen zu belegen. In der GUI ist beispielsweise ein Button zum Starten der Konvertierung enthalten, der eine Funktion des *BoilerplateHandler* aufruft.
- *AnnieHandler* - Der Annie-Handler realisiert den Zugriff auf die Bestandteile der Sprachverarbeitung aus dem Werkzeug GATE. Hauptsächlich betrifft dies die Verarbeitungslogik ANNIE, die konkrete Komponenten zur Textverarbeitung enthält (siehe Abschnitt 4.x). Im *AnnieHandler* werden dazu neben den Verarbeitungswerkzeugen auch die Textressourcen, die im Programm verarbeitet werden sollen, initialisiert. Diese *Language-Ressourcen* aus GATE beinhalten neben dem Text auch alle generierten Annotationen. Der *AnnieHandler* beinhaltet nach der Verarbeitung somit alle Informationen zu den Anforderungen in einer Liste. Auf Basis dieser Informationen kann dann die Transformation von Anforderungen in Boilerplates im *BoilerplateHandler* erfolgen.
- *BoilerplateHandler* - Die Klasse dient der Erzeugung von jeweils passenden Boilerplates zu einer Anforderung, indem aus dem Text alle relevanten Informationen zur Erzeugung passender Boilerplates extrahiert werden. Der *BoilerplateHandler* beinhaltet dazu sowohl Funktionen zur Identifikation passender Boilerplates auf Basis der Annotationen aus der GATE-Verarbeitung, als auch die Erzeugung von konkreten Boilerplate-Objekten aus der abstrakten Klasse *Boilerplate*, wenn die passenden Annotationen dazu gefunden wurden.
- *Boilerplate* - Die Klasse bündelt alle Kernmerkmale der Boilerplates, die durch das Programm erzeugt wurden. Dies betrifft in erster Linie

²Die Darstellung der GUI-Klasse in Abb. 4.x ist für eine bessere Übersichtlichkeit vereinfacht worden. Enthaltene Objekte und Funktionalitäten beziehen sich dabei jedoch nur auf die GUI-Elemente.

den Gesamttext, die einzelnen Satzbestandteile und Metainformationen über die Anforderung, wie etwa eine ID.

- *Boilerplate65c* und *Boilerplate85* - Wie bereits unter 3.x beschrieben, wurden im Zuge der Entwicklung dieses Werkzeugs zwei prototypische Boilerplates implementiert. Diese konkreten Boilerplate-Objekte beinhalten eine Repräsentation der GATE-Annotation einzelner Satzbestandteile als Zeichenketten, um daraus ganze Sätze zu erzeugen. Der Boilerplate-Handler "befüllt" somit konkrete Boilerplates mit allen erforderlichen Begriffen anhand der dort definierten Bestandteile.

- Die Klasse *Boilerplate65c* stellt eine Schablone für Systemfunktionen dar und enthält demnach Platzhalter für ein System und eine Funktion dieses Systems, die dann bei der Erzeugung konkreter Boilerplate-Objekte in das Grundgerüst des Satzes eingefügt werden. Typ 65c Boilerplates sehen demnach etwa so aus:

```
The complete system <<System>> shall perform
<<FunctionDescription>>.3
```

- Die Klasse *Boilerplate85* stellt vereinfacht eine Erweiterung der *Boilerplate65c* um eine Bedingung dar. Dementsprechend existiert auch hier ein System und eine Funktion, aber auch eine Bedingung, unter der die Funktion ausgeführt wird. Diese drei Bestandteile werden für die Erzeugung von Typ 85 Boilerplates benötigt. Das Grundgerüst der Sätze mit bedingten Funktionen sieht dabei wie folgt aus:

```
Under the condition: <<conditionText>> the actor:
<<conditionalActorText>> shall
<<functionDescriptionText>>.4
```

Die Funktionalität des Gesamtprogramms, das sich aus dem Zusammenspiel der KOMPONENTEN ergibt, ist in den folgenden Abschnitten dargestellt.

³«System» und «FunctionDescription» werden durch die Verarbeitung des Satzes in GATE eindeutig anhand passender Annotationen identifiziert und zugeordnet. Der Rest des Satzes ist fest hinterlegt.

⁴Eine Unterscheidung zu Typ65c scheint zunächst nicht trivial, da conditionText auch einfach leer sein kann und somit Typ 65c fälschlicherweise als Typ 85 erzeugt werden können. Durch die unterschiedliche ANnotation der beiden Typen schon bei der GATE-Verarbeitung kann dennoch eine verlässliche Unterscheidung erfolgen.

4.3 Werkzeug-Implementierung

In diesem Abschnitt werden wichtige Aspekte der praktischen Umsetzung des Programms genau erläutert. Die bezieht sich sowohl auf präzisere Vertiefung der Konzepte, also auch Quellcode-Beispiele der Implementierung zur Erläuterung des Programmvorgehens und die daraus resultierenden Interaktionsmöglichkeiten der Benutzer.

4.3.1 Sprachverarbeitung mit GATE

Dieser Abschnitt beschäftigt sich mit der automatischen Sprachverarbeitung der Anforderungsdokumente mit Hilfe des angekündigten NLP-Werkzeugs *GATE*. Dies beinhaltet, das Einlesen der Lastenhefte, die Erstellung einer Verarbeitungslogik, die Erstellung und Auswertung von Verarbeitungsregeln und das Einbinden von GATE in das spätere Programm. Fokus liegt dabei auf der praxisnahen Verwendung in dem Projekt. Ein genereller Überblick und Einordnung des Werkzeugs finden sich in der vorherigen Arbeit [RS18].

Einlesen von Lastenheften

In GATE benötigt man zum einlesen und verwenden von Dokumenten zunächst einen *Corpus*, also einen Textkörper. Dieser kann später mehrere Dokumente enthalten und diese verwalten. Außerdem wird eine Verarbeitungslogik immer auf einem Korpus (nicht auf einem Dokument) ausgeführt. Im Zusammenhang dieser Arbeit war es nicht nötig einen Corpus mit verschiedenen Dokumenten zu befüllen, daher ist zukünftig die Verarbeitung eines Corpus gleichbedeutend mit der Verarbeitung eines Dokuments.

Ein Corpus kann in der Benutzeroberfläche von GATE schnell als neue *language resource* (Sprachliche Ressource) erzeugt werden. In GATE Embedded ist dies ohne Benutzeroberfläche natürlich nicht möglich, lässt sich jedoch ähnlich einfach durch den Aufruf von *new Corpus* durchführen.

Sobald der Corpus erstellt ist, lassen sich dort Dokumente einfügen.⁵ Auch dies geschieht in GATE direkt über die Benutzeroberfläche. In GATE Embedded verwendet man hierzu den *add*-Aufruf eines Corpus-Objekts mit einem übergebenen Dokument.

⁵Alternativ kann auch erst ein Dokument eingelesen und anschließend ein Corpus um dieses angelegt werden.

4.3. WERKZEUG-IMPLEMENTIERUNG Analyse von Projektlastenheften

Sobald man einen Corpus mit mindestens einem Dokument hat, kann man auf diesem Corpus eine Verarbeitungslogik aufrufen (oft auch Verarbeitungs-Pipeline genannt, wegen der Aneinanderreihung mehrerer Verarbeitungskomponenten). Diese werden im folgenden Abschnitt genauer beschrieben.

GATE-Verarbeitungslogik

In GATE wird die sprachliche Analyse durch eine Verarbeitungslogik durchgeführt, welche aus verschiedene Verarbeitungsressourcen (*processing resource*) aufgebaut ist. Eine Standard-Verarbeitungslogik wird durch das GATE-Plugin *ANNIE* zur Verfügung gestellt. Die einzelnen ANNIE-Verarbeitungsressourcen sind in [RS18] beschrieben.

Führt man eine Verarbeitungslogik auf einem Corpus bzw. auf dessen Dokumenten aus, so werden durch die verschiedenen Verarbeitungsressourcen an bestimmten Stellen im Text bestimmte Informationen hinzugefügt. Dies bezeichnet man als *annotieren*. Dabei wird der Text an sich nicht geändert, die Annotationen können jedoch etwa durch farbige Unterlegung des betroffenen Texts in der GATE-Benutzeroberfläche gezeigt und als XML-Datei exportiert werden.

	ANNIE English Token...	ANNIE English Tokeniser
	ANNIE Gazetteer	ANNIE Gazetteer
	ANNIE Sentence Splitter	ANNIE Sentence Splitter
	ANNIE POS Tagger	ANNIE POS Tagger

Abbildung 4.3: Anforderung mit farbige markierten Token-Annotationen in GATE

Einer Annotation können sog. *Features* also Eigenschaften zugewiesen werden, welche dann noch weitere Informationen enthalten können, wie z.B. Länge der Annotation und die betroffene Zeichenkette.

4.3. WERKZEUG-IMPLEMENTIERUNG Analyse von Projektlastenheften

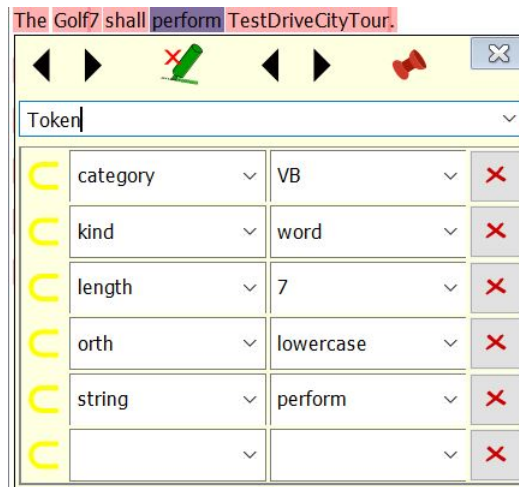


Abbildung 4.4: Ausgewählte Token-Annotation und dessen Features in GATE

Die für diese Arbeit verwendeten und besonders relevanten ANNIE- Verarbeitungsressourcen und die dadurch entstehenden Annotationen sind:

1. *ANNIE Tokenizer*: Der Tokenizer unterteilt den Text oberflächlich in Zeichen oder Zeichenketten, welche als Zahlen, Wörter (Groß- und Kleinschreibung wird unterschieden) und Interpunktion kategorisiert werden. Dies fügt an den betroffenen Stellen die *Token*-Annotation hinzu, welche jeweils Typ (Zahl, Wort usw.) und Länge in Zeichen . Die Verarbeitung durch den Tokenizer soll so effizient wie möglich sein und nur als Grundlage etwa für spätere Grammatikregeln dienen.
2. *ANNIE Gazetteer*: Ein Gazetteer beinhaltet relevante Eigennamen aus bestimmten Bereichen. Es handelt sich dabei um eine geschachtelte Liste, welche für jeden Bereich (beispielsweise Städte, Flughäfen, Fußball Clubs) über eine Liste der jeweils bekannten Eigennamen verfügt. Gazetteers werden verwendet, um domänenspezifisch Entitäten zu erkennen, da diese in einem normalen Wörterbuch nicht zu finden sind. Die vorgefertigten Listen sind also je nach Dokument individuell zu erweitern bzw. zu erstellen. Wird ein Listenelement im Text wiedererkannt, so erhält es Standardmäßig die *Lookup*-Annotation. Das lässt sich jedoch auch schon im Gazetteer ändern und jeweils Listen-abhängig bestimmen.

4.3. WERKZEUG-IMPLEMENTIERUNG Analyse von Projektlastenheften

3. *ANNIE Sentence Splitter*: Der Sentence-Splitter unterteilt den gesamten Text in einzelne Sätze. Dadurch wird jeder komplette Satz von einer *Sentence*-Annotation umschlossen. Dies wird später für den Part-of-Speech Tagger benötigt. Der Splitter verwendet dazu einen Gazetteer mit Abkürzungen, um Interpunktion wie etwa Fragezeichen, Ausrufezeichen und Punkte von anderen Sonderzeichen wie Semikola und Doppelpunkten zu unterscheiden.
4. *ANNIE POS Tagger*: Der Part-of-Speech Tagger annotiert alle Wörter und Symbole gemäß ihrer grammatikalischen Funktion. Hierfür wird keine neue Annotation angelegt, es werden jedoch das wichtige *category*-Feature an die Token-Annotationen angefügt. Dies ist wichtige Grundlage für grammatikalische Analyse von Sätzen. Dazu gehören dafür nutzt dieser Standard-Lexika und Regelwerke, welche auf dem Corpus-Training des Wall Street Journal beruhen.

	ANNIE English Tokeni...	ANNIE English Tokeniser
	ANNIE Gazetteer	ANNIE Gazetteer
	ANNIE Sentence Splitter	ANNIE Sentence Splitter
	ANNIE POS Tagger	ANNIE POS Tagger

Abbildung 4.5: Verwendete ANNIE-Verarbeitungsressourcen in GATE [eigene Darstellung]

JAPE Transducer

Die für die Verarbeitung wichtigsten Komponenten sind in der normalen ANNIE-Verarbeitungslogik jedoch noch nicht vorhanden. Die sogenannten *JAPE Transducer* werden individuell angelegt und verwenden die Annotationen der vorherigen Verarbeitungsressourcen, um dem Text wichtige Informationen hinzuzufügen. Diese können im Gegensatz zu den vorherigen jedoch auch von großer inhaltlicher bzw. semantischer Relevanz sein. Jape Transducer lassen sich wie die anderen Verarbeitungsressourcen hintereinander reihen und somit in eine Verarbeitungslogik einfügen.

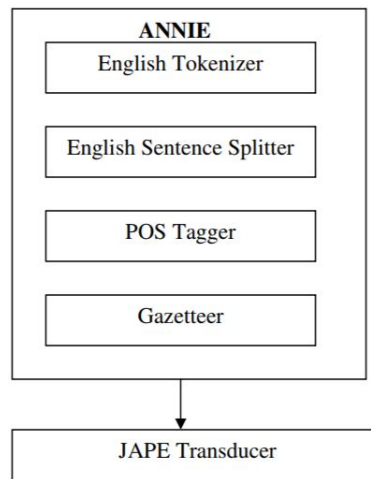


Abbildung 4.6: Verarbeitungslogik aus ANNIE und JAPE Transducern [TOL09]

In JAPE-Transducern⁶ wird die im Namen vorhandene *JAPE* (Java Annotation Patterns Engine) verwendet. Jape fungiert dabei als endlicher Zustandswandler⁷ der reguläre Ausdrücke prüft und abhängig davon Annotationen erzeugt oder JAVA-Code ausführt.

HIER EVTL QUELLE VERWENDEN

Um ein Verständnis von JAPE Transducern aufzubauen ist jedoch zunächst einer Erläuterung von JAPE-Dateien bzw. JAPE-Regeln nötig, welche im Folgenden gegeben ist.

Einführung zu JAPE-Regeln

Jape Transducer werden immer mit einer *Jape-Datei* initialisiert. JAPE-Dateien sind erkennbar an der *.jape*-Endung und können entweder eine JAPE-Regel oder eine JAPE-Hauptklasse (*main.jape*) darstellen.

Eine Hauptklasse wird dazu verwendet, mehrere JAPE-Regeln zu vereinen. Dadurch lassen sich auch Transducer mit mehreren Regeln initialisieren und unnötig lange Verarbeitungs-Pipelines verhindern.

⁶sowie in einigen der anderen Verarbeitungsressourcen

⁷Für Erläuterung siehe [RS18]

Die JAPE-Regeln bilden die eigentliche Logik der jeweiligen Verarbeitungsressource und ähneln kleinen Programmen mit der Erkennung einer Eingabe und der Erzeugung einer bestimmten Ausgabe. Daher wird im Folgenden auch von JAPE-Code gesprochen.

Der Code in einer JAPE-Regel ist immer unterteilt in einen Regel-Kopf eine linke Seite (*LHS* für left hand side) und eine rechte Seite (*RHS* für right hand side). Diese Benennung kann jedoch irreführend sein, wenn man eine JAPE-Regel betrachtet, da die Regeln von oben (Kopf, LHS) nach unten (RHS) aufgebaut sind, so wie es für Quellcode üblich ist. Abbildung X.X zeigt beispielhaft eine stark vereinfachte JAPE-Regel.

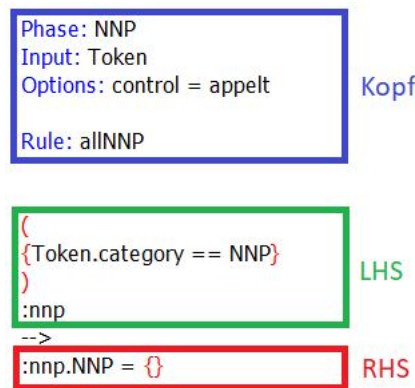


Abbildung 4.7: Beispielhafte JAPE-Regel [eigene Darstellung]

Im Kopf der Regel befinden sich die Parameter *Phase*, *Input*, *Options* und *Rule*.

1. **Phase:** Hier lässt sich eine Phase festlegen, in welcher die Regel ausgeführt werden soll. Dies kann wichtig sein, wenn z.B. eine Regel eine Annotation verwenden möchte, welche zuvor von einer anderen Regel angelegt werden muss. In unserem Projekt wurden für eine bessere Übersicht verschiedene Transducer angelegt, welche dann hintereinander durchlaufen werden. Daher wird die Phase im weiteren Verlauf der Arbeit nicht weiter verwendet oder behandelt.
2. **Input:** Das Definieren des Inputs ist vor allem später für die LHS von entscheidender Bedeutung. Hier wird angegeben, welche Annotationen in den regulären Ausdrücken abgefragt werden können.

Wenn ein Annotationstyp im Input nicht angegeben ist, jedoch in der LHS verwendet wird, führt dies zu einem Fehler beim Erstellen des Transducers. Wenn im Input Annotationen angegeben sind, welche in der LHS nicht verwendet werden, schadet dies unnötig der Laufzeit des Transducers.

3. Options: Hier können verschiedene Optionen zum Verhalten der Regel in bestimmten Fällen angegeben werden. Dies beinhaltet hauptsächlich die Anpassung der *control*-Option. Diese entscheidet über das Resultat der Regel, wenn sich etwa mehrere Annotationen der Ausgabe überschneiden. So kann etwa jede einzelne Annotation oder auch nur die längste Annotation (welche die kleineren abdeckt) ausgegeben werden. Andere Optionen wie *negationGrouping* können benötigt werden, um bestimmte reguläre Ausdrücke in der LHS zu ermöglichen.
4. Rule: Hier wird der Name der Regel definiert. Dieser weist üblicherweise auf die Funktion der Regel, bzw. die erzeugte Annotation hin. Außerdem wird er oftmals als Feature der erzeugten Annotationen hinzugefügt, um die Nachvollziehbarkeit einer Verarbeitungslogik zu erhöhen.

In der JAPE-Regel befindet sich unterhalb des Kopfes die LHS. Diese umfasst eine Menge von regulären Ausdrücken, basierend auf den Annotationen, welche als Input deklariert wurden. Diese werden in dem Text gesucht und auf die angegebenen Bedingungen geprüft. Sollten gefundene Annotationen diese erfüllen, so werden diese mit einem Label versehen.

Ein Pfeil trennt die LHS von der RHS. Unterhalb der Pfeile befindet sich die RHS, in welcher abhängig von den vergebenen Labels nun neue Annotationen und deren Features mit Werten definiert werden können.

Alternativ kann sich auf der RHS in einer JAPE-Regel auch JAVA-Code befinden, um erweiterte Funktionalitäten nutzen zu können. Dies war für das Projekt jedoch nicht nötig, da die Ergebnisse der JAPE-Regeln ohnehin später in ein JAVA-Programm verwendet wurden. Daher wird diese Option im Verlauf der Arbeit nicht weiter berücksichtigt.

Der Aufbau von JAPE-Regeln soll an dem Beispiel in Abbildung X.X genauer erläutert werden:

Zunächst wird die Phase *NNP* benannt und die *Token*-Annotation als Input deklariert. Die *control*-Option wird auf *appelt* gesetzt. Diese wird im

4.3. WERKZEUG-IMPLEMENTIERUNG Analyse von Projektlastenheften

Normalfall verwendet und bewirkt, dass eine betroffene Stelle im Text nur von einer Regel genutzt werden kann⁸. Die Regel selbst wird mit *allNNP* benannt, dies wird jedoch im Rest der Regel nicht wieder verwendet.

In der LHS wird mit der Abfrage

$$\{Token.category == NNP\}$$

das *category*-Feature der *Token*-Annotation auf den Wert *NNP* geprüft⁹. Die Abfrage entspricht also der Form

$$\{Annotation.Feature == Wert\}$$

Alle Tokens, welche diese Bedingung erfüllen, werden daraufhin mit dem Label *nnp* versehen, welches auf die RHS der Regel übernommen wird.

Unterhalb des Pfeils wird nun durch die Zeile

$$: nnp.NNP = \{\}$$

jede mit *nnp* markierte Stelle mit der *NNP*-Annotation versehen. In den geschweiften Klammern könnten dieser Annotation außerdem Features übergeben werden. Möchte man wie zuvor erwähnt den Namen der Regel als Feature hinzufügen, so lässt sich dies mit der folgenden Ergänzung tun:

$$: nnp.NNP = \{rule = "allNNP"\}$$

Damit ist eine neue Annotation Namens *NNP* mit dem Feature *rule* erzeugt, welche von späteren Regeln mit der einfachen Abfrage $\{NNP\}$ genutzt werden kann, ohne den Wert eines Features abfragen zu müssen.

⁸Setzen zwei Regeln am gleichen Startpunkt an, so erhält die Regel mit der längeren Übereinstimmung Vorrang, sind beide gleich lang, so erhält die Regel mit der höheren Priorität Vorrang, haben beide die gleiche Priorität, so erhält die Regel Vorrang, welche als erste in der Grammatik definiert wurde.[TOL09]

⁹Erinnerung: Die *Token*-Annotation wird von dem Tokenizer erstellt, das *category*-Feature und dessen Wert jedoch von dem POS-Tagger

JAPE-Regeln in der Praxis

In diesem Abschnitt werden einige der in der Sprachverarbeitung verwendeten JAPE-Regeln und deren Funktionsweise genauer erläutert.

Ursprünglich war der Ansatz für die Erkennung von Systemnamen eine genaue grammatikalische Analyse zu nutzen um solche Namen komplett erfassen zu können. Dieser wurde jedoch nach verschiedenen Tests für zu unpräzise befunden. Grund dafür war, dass in den Systemnamen keine ausreichende Regelmäßigkeit erkannt werden konnte, um eine verlässliche Erfassung zu gewährleisten.

Daher wurde als nächstes ein Ansatz mit einer simpleren, jedoch auch besser umsetzbaren Methode verfolgt. Durch die Nutzung eines Gazetteers, kann die korrekte Erkennung von Eigennamen stark verbessert werden. Der einzige Nachteil ist, dass diese Eigennamen zuvor bekannt sein müssen, dies ist jedoch im betrieblichen Kontext gut umsetzbar. Um dies zu demonstrieren wurde ein ANNIE-Gazetteer mit einigen Systemnamen angelegt und sowohl für Beispiele, als auch für komplette Lastenhefte getestet. Daher wird im Folgenden anstelle einer JAPE-Regel der verwendete Gazetteer vorgestellt. Solch ein Gazetteer besteht aus einer Menge von *.lst*-Dateien welche als Listen fungieren und mindestens einer *.def*-Datei, welche definiert wie mit den einzelnen Listen umzugehen ist. In Abbildung X.X ist ein Ausschnitt einer solchen Listendatei zu sehen. Abbildung X.X zeigt die direkte Definition einer bestimmten Annotation innerhalb der Definitionsdatei. Dies ermöglicht das Einsparen einer JAPE-Regel, welche die standardisierte *Lookup*-Annotation abfragen und einordnen muss.

```
TIM
RKE
LED driver
RF transmitter and Receiver
BCM
Remote
receiver
transponder
```

Abbildung 4.8: Ausschnitt der *SystemName*-Liste des Annie-Gazetteers [eigene Darstellung]

4.3. WERKZEUG-IMPLEMENTIERUNG Analyse von Projektlastenheften

```
complete_system.lst:::CompleteSystem
module.lst:::Module
```

Abbildung 4.9: Definitionsdatei des Annie-Gazetteers [eigene Darstellung]

Die Regel *recognize_conditional_actor* wird dazu verwendet, im Falle einer Anforderung mit Bedingung den Akteur zu ermitteln, welcher durch die Bedingung eingeschränkt wird. Diese Anforderungen können dann später in die *Boilerplate85* überführt werden, welche in Abschnitt X.X dargestellt ist.

```
Phase: ConditionalRequirement
Input: Token Split Condition
Options: control = appelt negationGrouping=false

Rule: ConditionalRequirement

({Condition})?
({Token.string==","})?
({Token.category=="DT"})?

(
  (
    {Token, !Split, !Token.string=="shall", !Token.string=="should"})*
  ):conditionalactor

  ({Token.string=="shall"} | {Token.string=="should"})

)

-->
:conditionalactor.ConditionalActor = {}
```

Abbildung 4.10: JAPE-Regel zur Erkennung von eingeschränkten Akteuren [eigene Darstellung]

In der Phase und der Benennung der Regel sind soweit keine Besonderheiten. Als Input sind hier jedoch *Token*, *Split* und *Condition*. Token ist die bereits bekannte Annotation des Tokenizers und Split ist eine Annotation des Sentence Splitters, welche das Ende eines Satzes (z.B. durch Punkt oder Leerzeile) markiert. Außerdem wird die eigens erstellte Condition-Annotation verwendet, welche auf Grundlage von Schlagwörtern wie *if* oder *when* die Bedingung in einem Satz erkennen und markieren kann¹⁰. Außerdem wird die

¹⁰Auch zu dieser Annotation gibt es natürlich eine Regel, diese wird jedoch aus Platz-

Option *negationGrouping* auf *false* gesetzt, um es zu ermöglichen mehrere negierte Abfragen auf der gleichen Token-Annotation auszuführen.

Generell muss die Regel die folgenden zwei Satzstellungen berücksichtigen können:

1. Unter der Bedingung *A*, soll das System *B* die Aktion *C* durchführen.
2. Das System *B* soll die Aktion *C* durchführen, unter der Bedingung *A*.

Die LHS der Regel beginnt daher mit einer optionalen Abfrage auf eine Condition, ein Komma und einen *Determiner* (the, this, a usw.). Dies sorgt dafür, dass diese Elemente nicht mit als *ConditionalActor* markiert werden, falls sie wie in 1. vor dem Akteur im Satz auftauchen.

Daraufhin werden die folgenden Tokens als *conditionalactor* getaggt, jedoch nur solange:

1. keine Split-Annotation enthalten ist,
2. kein *shall*
3. und kein *should* im Text auftauchen.

Nachlaufend ist es jedoch nötig, dass eines dieser Wörter auftaucht (außerhalb des *conditionaleactor*-Taggings), da der Satz sonst nicht für die gewollte Boilerplate geeignet sein kann.

Damit ist LHS bzw. die Erkennung abgeschlossen und die als *conditionalactor* getaggtten Bereiche werden mit der *ConditionalActor*-Annotation versehen. Features werden dabei nicht benötigt.

Sobald alle benötigten Satzbestandteile erfasst sind, lässt sich auf deren Grundlage nun eine geeignete Boilerplate definieren. Diese wird mit der JAPE-Regel *BP85_bsp* in Abbildung X.X erkannt und markiert.

JAPE Transducer in GATE

Im Regelkopf ist hierbei zu beachten, dass alle für die Boilerplate benötigten Satzbestandteile als Input definiert werden. In diesem Fall werden also die

gründen nicht genauer erläutert und befindet sich mit den restlichen Regeln in den Anlagen der Arbeit

4.3. WERKZEUG-IMPLEMENTIERUNG Analyse von Projektlastenheften

```
Phase: BoilerPlateConversion
Input: Token Split Condition ConditionalActor FunctionDescription
Options: control = appelle debug = true

Rule: BP85_bsp
(
  ({Token, !Split}) *
  ({Condition})
  ({Token, !Split}) *
  ({ConditionalActor})
  ({Token, !Split}) *
  ({FunctionDescription})
  ({Token, !Split}) *

  |

  ({Token, !Split}) *
  ({ConditionalActor})
  ({Token, !Split}) *
  ({FunctionDescription})
  ({Token, !Split}) *
  ({Condition})
  ({Token, !Split}) *

)
:ann

-->
:ann.Boilerplate85 = {rule=BP85_bsp,
text=:ann@cleanString,
conditionText=:ann.Condition@cleanString,
conditionalActorText=:ann.ConditionalActor@cleanString,
functionDescriptionText=:ann.FunctionDescription@cleanString
}
```

Abbildung 4.11: JAPE-Regel zur Markierung von Sätzen für Boilerplate85 [eigene Darstellung]

Condition-, *ConditionalActor*- und *FunctionDescription*-Annotationen verwendet.

Daraufhin werden in den regulären Ausdrücken der LHS die zwei Satz-schemata abgefragt, welche schon bei der Erkennung des *ConditionalActor* eine wichtige Rolle gespielt hatten:

1. Unter der Bedingung *A*, soll das System *B* die Aktion *C* durchführen.
2. Das System *B* soll die Aktion *C* durchführen, unter der Bedingung *A*.

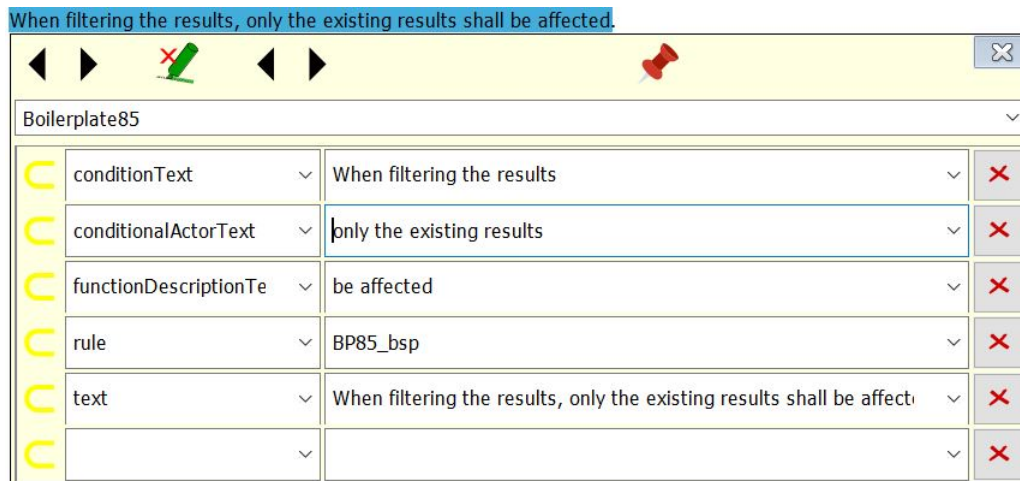
Zwischen den Satzbestandteilen vorkommende Tokens werden dabei ignoriert, solange diese keine *Split*-Annotation enthalten, also mit Punkten oder Zeilenumbrüchen den Satz beenden. Alle Sätze die mit den Satz-schemata

4.3. WERKZEUG-IMPLEMENTIERUNG Analyse von Projektlastenheften

entsprechen und die gestellten Bedingungen erfüllen werden als *ann* gelabelt und auf der RHS der Regel verarbeitet.

Im Gegensatz zu den vorherigen Regeln ist auf dieser RHS die Vergabe von Features von entscheidender Bedeutung. Da man in dieser Regel verschiedene Annotationen zusammenfasst, ist es wichtig Informationsverlust zu verhindern, vor allem, weil die Satzbestandteile später noch vorliegen müssen um die Boilerplates zu befüllen. Daher erhält jede erstellte Boilerplate-Annotation auch ihren kompletten Text, sowie jeweils den Text jedes wichtigen Satzbestandteils als Feature. Dies macht es außerdem später im Programm einfacher diese abzufragen. Alle Features der Annotation sind in Abbildung X.X zu sehen.

When filtering the results, only the existing results shall be affected.



Boilerplate85			
C	conditionText	When filtering the results	X
C	conditionalActorText	only the existing results	X
C	functionDescriptionTe	be affected	X
C	rule	BP85_bsp	X
C	text	When filtering the results, only the existing results shall be affect	X
C			X

Abbildung 4.12: Alle Features eines *Boilerplate85*-Satzes in GATE [eigene Darstellung]

JAPE Transducer in GATE

Sobald alle benötigten JAPE-Regeln definiert sind, lassen sich diese unter JAPE-Hauptdateien zusammenfassen um die benötigte Anzahl von JAPE Transducern zu verringern. Es sollte jedoch nur soweit zusammengefasst werden, dass einzelne Transducer noch austauschbar sind, wenn man die Sprachverarbeitung ändern möchte. Die Transducer lassen sich (mit dem ANNIE-Plugin) über die Benutzeroberfläche von GATE leicht als Verarbeitungsressource anlegen und erhalten beim Initialisieren direkt die jeweilige

JAPE-Datei. Syntaxfehler innerhalb der Dateien werden sofort erkannt und im Hauptfenster ausgegeben. Sollten keine Fehler vorhanden sein erhält man die fertige Verarbeitungsressource, welche man an beliebiger Stelle in die Verarbeitungs-Pipeline einbauen kann. Obwohl in dem Programm GATE Embedded verwendet wurde und die Benutzeroberfläche somit nicht direkt verwendet werden konnte, bietet GATE die wichtige Funktion Corpora und Verarbeitungslogiken zu speichern und zu laden. Somit kann in man eine Verarbeitungslogik in GATE erstellen diese als sogenannte *.gapp*-Datei speichern und anschließend in der Embedded-Version von GATE innerhalb des Programms laden. Außerdem lässt sich dabei der verwendete Corpus leicht ändern bzw. neu bestimmen. Das Laden einer solchen Datei in das Programm wird im folgenden Abschnitt genauer erläutert.

4.3.2 Converter-Funktionalität

In diesem Abschnitt wird die Funktionalität der Converter-Komponente des Programms erläutert. Diese dient, wie bereits beschrieben, zur automatisierten Transformation der gelesenen Lastenhefte in sog. Boilerplates. In unserem Programm greift die Verarbeitungslogik dabei auf die Funktionen der unter 4.x beschriebenen Klassen bzw. Methoden, sowie auf die im vorigen Abschnitt behandelten JAPE-Regeln zur Textannotation zu.

Diese Ressourcen werden im Programm dann nacheinander genutzt, um die Boilerplate-Konvertierung durchzuführen. In der Regel sind verschiedene Programmfunktionen dabei zur Nutzerinteraktion Buttons in der GUI zugeordnet (siehe Abs. 4.x). Für ein besseres Verständnis der Zusammenhänge ist das R2BC-Klassendiagramm, sowie der kommentierte Quellcode des Prototypen im Anhang hilfreich.

Listen-Architektur

Der Convert-Vorgang teilt sich auf die zwei Klassen AnnieHandler und BoilerplateHandler auf. AnnieHandler dient dabei zur Erstellung von Annotationen über den Text durch die Anbindung an GATE. BoilerplateHandler erstellt auf Basis der dort gefundenen Annotationen dann entsprechende Boilerplates.

Es ist dabei zu beachten, dass mehrere Deutungen bzw. erzeugbare Boilerplates aus einem Satz existieren können. Ferner muss vor dem Export des

konvertierten Textes in ein Ausgabedokument die eine richtige Konvertierung jedes Satzes gespeichert werden.

Das Datenmodell des Programms muss daher sowohl den ursprünglichen Text mitsamt Annotationen beinhalten, als auch alle möglichen Konvertierungen zu einem Satz in Boilerplates auflisten können.

Dieser Ansatz hat zur konkreten Erstellung von drei Listen im Programm geführt, *sentenceList*, *bpList* und *bpExportList*. Die Listen sind in Abb. 4.x prinzipiell in ihrer Verknüpfung dargestellt.

LISTENDARSTELLUNG Datenmodell R2BC aus drei Listen zur Speicherung von Annotationen(1), möglichen Konvertierungen(2) und Export(3) [eigene Darstellung]

- *sentenceList* repräsentiert dabei alle Paragraphen des Quelldokuments, da Anforderungen immer durch einen Zeilenumbruch getrennt sind.¹¹ soll dabei gefundene GATE-Annotationen zu jedem Satz beinhalten, die mithilfe der Klasse AnnieHandler angebundene Verarbeitung generiert werden.
- Auf dieser Basis erfolgt dann die Befüllung der *bpList* mit möglichen Boilerplate-Konvertierungen zu jedem Satz in der Klasse BoilerplateHandler.¹² Die Annotationen dienen dabei zur Erzeugung der Boilerplates für die *bpList*.
- In der *bpExportList* sind alle Sätze enthalten, die nach dem Programm-durchlauf exportiert werden sollen. Nach der Wahl der zu speichernden Konvertierung durch den Nutzer wird der entsprechende Satz hier eingetragen.

sentenceList und *bpList* sind dementsprechend zweidimensional und beinhalten jeweils Annotationen (*sentenceList*, horizontal) und Boilerplate-Konvertierungen (*bpList*, horizontal) zu jedem Satz des Lastenheftes (vertikal). Durch den Aufruf verschiedener Methoden der Klasse BoilerplateHandler werden beide Liste automatisch initialisiert und befüllt.

Anzumerken ist, dass *bpExportList* prinzipiell auch automatisch befüllt werden kann, wenn etwa keine Konvertierung gefunden und somit der ursprüngliche Satz beibehalten werden soll. Auch könnte bei hinreichender

¹¹In unserem Programm wird konkret die Sentence-Annotation aus dem ANNIE-SentenceSplitter von GATE verwendet, um Sätze zu trennen.

¹²Wie bereits gezeigt, wurden dazu verschiedene verschachtelte JAPE-Regeln entwickelt, eine mögliche BP-Konvertierung signalisieren bzw. annotieren.

Genauigkeit gleich die bestmögliche Konvertierung des Satzes gespeichert werden. Im Regelfall wird die Konvertierung jedoch durch den Nutzer im Programm ausgewählt bzw. bestätigt. Im folgende

AnnieHandler

Die Klasse *AnnieHandler* stellt den ersten Schritt bei der Verarbeitung der Lastenhefte dar. Die Sätze aus dem Lastenheft müssen vor der tatsächlichen Konvertierung in Boilerplates mithilfe des NLP-Werkzeugs GATE annotiert werden. Dazu beinhaltet der *AnnieHandler* verschiedene Objekte zur Verwaltung der GATE-Ressourcen.

Dazu dienen ein *Document* für das eingelesene Dokument, sowie ein *Corpus* und ein *CorpusController*, der auf die ANNIE-Module in GATE-Embedded zur Verarbeitung zugreift.

Zunächst wird in der Methode *initFromRestore()* (Abb.4.x) die ANNIE-Pipeline initialisiert, indem eine GATE-Sitzung aus einer *.gapp*-Datei wiederhergestellt wird. Diese Datei wurde vorher angelegt, um einen direkten Zugriff auf die dort hinterlegten Module aus dem R2BC zu ermöglichen.

```
// initialise the GATE library
Gate.init();

//restore from File
controller = (CorpusController)
PersistenceManager.loadObjectFromFile(new File("ANNIE.gapp"));
```

Abbildung 4.13: Initialisierung der GATE-Ressourcen in *initFromRestore()*-Methode der Klasse *AnnieHandler* [eigene Darstellung]

Nachdem nun ein Dokument eingelesen wurde, kann dies dem *Document*-Object zugeordnet und auf dieser Basis der *Corpus* zur Speicherung der Annotationen über das Lastenheft initialisiert werden. In diesen *Corpus* werden während der Verarbeitung Annotationen über den Text gesammelt. Für den weiteren Ablauf betrifft dies insbesondere gefundene Boilerplate-Annotationen, sowie die dafür nötigen Satzbestandteile, die gesammelt werden sollen.

Dazu kann nun mit der Verarbeitung durch die ANNIE-Pipeline begonnen werden. Die Methode *execute()* führt dazu den *CorpusController*, der die konkrete ANNIE-Pipeline für den R2BC enthält, auf dem *Corpus* aus (Abb.

4.x). Der AnnieHandler beinhaltet dazu die Methode *convert()*, die in Abb. 4.x abgebildet ist.

```
//create Corpus for document from file system
corpus = Factory.newCorpus("MyCorpus");
controller.setCorpus(corpus);
File source = Main.inputFile;
doc = Factory.newDocument(source.toURI().toURL());
corpus.add(doc);

controller.execute();

bpHandler = new BoilerplateHandler(this);

//Load paragraphs and initialize Lists
bpHandler.initializeParagraphs();

//Traverse BP-Annotations and add to sentenceList
bpHandler.findPossibleConversions();

//convert annotated sentences to matching BPs, store Boilerplates
bpHandler.convertBPs();
```

Abbildung 4.14: Ausführung der ANNIE-Pipeline, Verarbeitung der Annotationen in BoilerplateHandler in *convert()*-Methode der Klasse AnnieHandler [eigene Darstellung]

Nach der Ausführung der GATE-Verarbeitung liegen die gefundenen Annotationen innerhalb des *Corpus*-Objektes vor, wodurch mit der weiteren Verarbeitung im BoilerplateHandler begonnen werden kann. Dazu wird mithilfe von *new* ein neuer BoilerplateHandler erzeugt. In der Methode *convert()* (Abb. 4.x) findet daher auch direkt der Aufruf verschiedener Methoden der Klasse BoilerplateHandler statt, der die Konvertierung der gefundenen Sätze in Boilerplates und die Initialisierung der Listen aus dem vorigen Abschnitt durchführt.

BoilerplateHandler

Im ersten Schritt werden dazu der Text satzweise getrennt. Dies erfolgt unter Nutzung der sentence-Annotation aus GATE. Dies bietet den Vorteil, dass nun ein satzweiser Zugriff auf Annotationen besteht.

4.3.3 GUI - Nutzerinteraktion und Workflow

Für ein besseres Verständnis wird sich hier auf ein kleines Beispiel beschränkt, ausführliche Tests mit realen Lastenheften aus dem betrieblichen Alltag werden unter 6.x behandelt. Die Implementierung wird daher anhand der Konvertierung der folgenden Sätze demonstriert, die in einem eingelesenen Lastenheft enthalten sein könnten:

The Golf7 shall perform TestDriveCityTour.

The Golf7 shall perform Charge4FreeCharging.

The EBus2X is part of Golf7.

The durability of EBus2X shall be at least 12000 h.

If there is a connection to the internet, the Bluetooth connected phone shall transmit the data to a server.

When filtering the results, only the existing results shall be affected.

Dieses prototypische Lastenheft wird im Laufe dieses Abschnittes von natürlicher Sprache in formale Boilerplates konvertiert.

4.4 Mögliche Erweiterungen

In diesem Abschnitt werden mögliche Verbesserungen bzw. Erweiterungen des Programms aufgelistet und erläutert. Diese konnten hauptsächlich auf Zeit- bzw. Aufwandsgründen nicht während der Arbeit bei HELLA implementiert werden. Einige erfordern außerdem ein hohes Maß an spezifischer Fachkenntnis in bestimmten Bereichen.

Die Erweiterungen sind zur Übersichtlichkeit in drei wichtige Bereiche unterteilt:

1. Benutzeroberfläche: Hier werden die Erweiterungen dargestellt, welche im direkten Zusammenhang mit der Nutzerinteraktion des Programms stehen und neue Ansichten auf der Benutzeroberfläche ermöglichen.

2. Nutzbarkeit: Hierzu gehören alle Erweiterungen, welche die Verwendbarkeit des Programms verbessern, jedoch nicht von Nutzer gesteuert werden. Dazu gehören Aspekte die Performanz, Abdeckungsgrad und präzise Sprachverarbeitung.
3. Konzept: Hier sind alle Erweiterungen aufgelistet, welche das Programm konzeptuell Verändern, wichtige Vorarbeiten leisten oder alternative Technologien verwenden.

4.4.1 Benutzeroberfläche

Im Folgenden sind die Erweiterungen zur Verbesserung der Benutzeroberfläche und dessen Funktionalität aufgeführt.

Kontext-Anzeige

Eine mögliche Verbesserung der Benutzeroberfläche könnte durch das Anzeigen von relevantem Kontext zu der aktuellen Anforderung erzielt werden. Dies bedeutet, dass der Nutzer die aktuelle Anforderung besser einordnen kann, wenn die beispielsweise mit der vorherigen oder der darauffolgenden zu tun hat.

Wie weit der relevante Kontext ausfällt ist hierbei an konkreten Lastenheften zu testen. Mögliche Ausführungen wären etwa das Anzeigen der umgebenden Anforderungen oder der letzten Information vor der Anforderung.

Außerdem könnte dies von entscheidender Bedeutung bei der Auflösung von Referenzen (wie zum Beispiel durch Pronomen) sein. Wörter wie „it“, „that“ usw. sollten zwar in Lastenheften wegen der Entstehung von Uneindeutigkeiten nicht verwendet werden, jedoch kann man dies in der Praxis nicht annehmen. Daher wären die Markierung von betroffenen Wörtern oder gar deren automatische Auflösung sehr nützlich.

Markieren von Unterschieden

Eine wichtige Hilfestellung bei der Übersichtlichkeit und somit bei der Verarbeitungsgeschwindigkeit durch den Programm-Nutzer könnte durch die farbliche Markierung von relevanten Textstellen erzeugt werden.

Wie in Abbildung X.X gezeigt, werden der Ursprungssatz und mögliche Übersetzungen in Boilerplates auf dem selben Nutzer-Bildschirm abgebildet.

Jedoch kann es durch die vorgegebene Satzstruktur der Boilerplates zur Umstellung der Eingabesatzes kommen. Dies ist für den Nutzer nur schwer auf den ersten Blick zu erkennen.

Eine farbige Markierung der wichtigen Satzbestandteile im Eingabesatz und der entsprechenden Gegenstücke in der übersetzten Boilerplate könnten dort für einen sofortigen Überblick sorgen. Die Markierung könnte abhängig von den durch GATE vergebenen Annotationen stattfinden und würde es dem Nutzer ermöglichen einige der angezeigten Möglichkeiten direkt auszuschließen bzw. zu bestätigen.

Vorher/Nachher Abgleich

Falls der Nutzer beabsichtigt die Ergebnisse der Dokumentkonvertierung noch einmal zu überprüfen könnte eine eingebaute Ansicht sehr nützlich sein. Diese zeigt das Eingabe- und das Ausgabedokument nebeneinander an. Gegebenenfalls könnten die Eingabe-Anforderungen passen zu den entsprechenden Übersetzungen markiert oder gefärbt werden.

Dadurch würden Fehler durch den Nutzer noch weiter eingeschränkt und somit die Konsistenz des Ausgabedokuments stark verbessert. Vor allem für die spätere Weiterverarbeitung durch den Delta-Analyser ist dies wichtig.

Eingabe des Dokumenttyps

Eine weitere Möglichkeit für den Nutzer das Programm zu verbessern könnte darin bestehen die Art des Eingabedokuments zu spezifizieren. Da in der Automobilindustrie viele verschiedene OEM mit verschiedenen Standards Pflichtenhefte für die Zulieferer verfassen, sind die eingehenden Anforderungsdokumente sehr heterogen und können auch bei der Formatierung voneinander abweichen.

Wenn der Nutzer jedoch die Möglichkeit hat, z.B. den OEM schon vor der Verarbeitung einzugeben und das Programm dementsprechend dynamisch angepasst wird, könnten auch ohne großen Aufwand für den Nutzer Lastenhefte verschiedener OEMs verarbeitet werden.

4.4.2 Nutzbarkeit

Hier werden die Erweiterungen behandelt, welche sich positiv auf die Nutzbarkeit des Programms auswirken könnten. Dazu gehören vor allem Performanz-

und Automatisierungsaspekte.

Implementierung weiterer Boilerplates

Eine naheliegende Verbesserung des Programms wäre das umsetzen weiterer Boilerplates. Dies ist jedoch nicht nur wichtig, um die Abdeckung des Testes zu verbessern, sondern auch um die Auswahl an Boilerplates zu verbessern.

Zunächst ergibt sich durch das Implementieren weiterer Boilerplates die Möglichkeit mehr Anforderungen des Quelldokuments in Boilerplates zu überführen. Damit geht jedoch auch ein Problem einher. Bei mehr Boilerplates steigt auch die Anzahl der vorgeschlagenen Boilerplates zu den einzelnen Sätzen. Im Optimalfall wird jedoch nur genau eine Boilerplate als Übersetzung vorgeschlagen, und zwar die richtige.

Es gilt also einen Anstieg an unpassend vorgeschlagenen Boilerplates so gering wie möglich zu halten und somit die Nutzbarkeit des Programms zu gewährleisten. Dies lässt sich erreichen in dem man die hinzugefügten Boilerplates abhängig von den verwendeten Satzbestandteilen in einer Hierarchie anordnet.

Angenommen eine Boilerplate *A* enthält die Satzbestandteile *1*, *2* und *3* und die Boilerplate *B* enthält *1*, *2*, *3* und *4*. Ein Satz der die Bestandteile *1-4* beinhaltet würde von in beide Boilerplates konvertierbar sein, da die Boilerplate *A* jedoch Bestandteil *4* nicht berücksichtigen kann, führt diese sicherlich zu Informationsverlust und ist somit falsch. Es lässt sich also schließen, dass die Boilerplate *A* niemals richtig sein kann, sobald Boilerplate *B* eine mögliche Übersetzung darstellt. Somit muss auch Boilerplate *A* nicht als möglicher Vorschlag angegeben werden.

JAPE-Optimierung

Die in der Sprachverarbeitung implementierten JAPE-Regeln wurden lediglich im Hinblick auf Funktionalität entworfen. Aufgrund der generell niedrigen Laufzeit¹³ wurde die Performanz der Regeln nicht extra optimiert. Es ist also möglich, dass dort noch Verbesserungspotential ausgeschöpft werden kann.

¹³Die Testergebnisse zur Laufzeit des Programms werden in Abschnitt X.X genauer erläutert

Die Implementierung weiterer Boilerplates sollte die Laufzeit des Programms nicht sonderlich beeinflussen, da der mit Abstand größte Annotations-Aufwand bei dem Tokenizer und dem Part Of Speech Tagger liegt. Diese werden jedoch auf jeden Fall genau ein mal ausgeführt.

Außerdem ist es möglich die JAPE-Transducer durch sogenannte *JAPE+*-Transducer zu ersetzen. Diese sind vor allem bei der Verarbeitung langer Dokumente sehr viel effizienter als herkömmliche JAPE-Transducer. Daher sind sie in dem Anwendungsfall mit Anforderungsdokumenten besonders geeignet. Abbildung X.X zeigt Testergebnisse der GATE-Entwickler zu den verschiedenen Transducern [GM19].

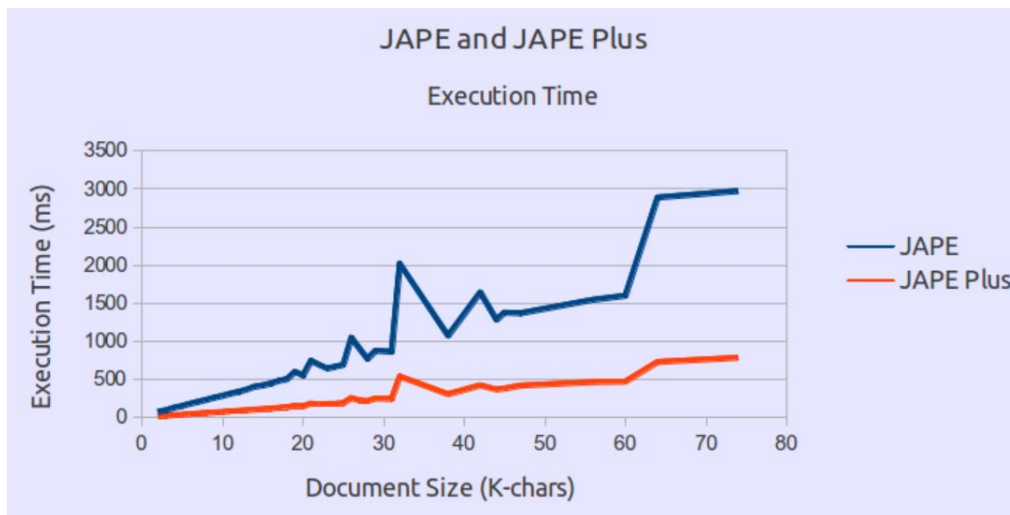


Abbildung 4.15: JAPE und JAPE+ Zeitaufwand abhängig von der Dokumentlänge [GM19]

Größtmögliche Automatisierung

Filtern von Anforderungen

4.5 Test

4.5.1 Methodik

4.5.2 Durchführung

4.5.3 Ergebnisse

Kapitel 5

Delta-Analyse

5.1 Architektur

5.2 Implementierung

5.3 mögliche Erweiterungen

5.4 Test

5.4.1 Methodik

5.4.2 Durchführung

5.4.3 Ergebnisse

Kapitel 6

Evaluation

6.0.1 Umsetzung der NLP-Architektur in unserem Werkzeug

In diesem Abschnitt wird die Einordnung unseres Prototypen als NLP-Werkzeug in die von [COP04] definierte Architektur behandelt. Die verschiedenen Schritte aus Kap. 2.2. haben wir bei der Analyse und Verarbeitung der Lastenhefte in Programmbestandteilen umgesetzt. Anhand der Verarbeitungskette unseres Werkzeugs können wir die prinzipiellen Aufgaben von NLP-Werkzeugen dort wie folgt einordnen:

1. *Input Processing* - Einlesen von Quelldaten: Zunächst kann der Nutzer Dateien für die Verarbeitung im Programm eingeben. Dazu wählt er diese Dateien in der GUI des Programms aus. Unser Werkzeug extrahiert daraus dann den enthaltenen Text.
2. *Morphologische Analyse* - Annotation der Worttypen: Mithilfe der Analyse des Textes durch die GATE-Verarbeitungslogik ANNIE werden einzelne Wörter aus dem Text auf ihre sprachliche Funktion hin überprüft. Dies ermöglicht die weitere Untersuchung bestimmter Worttypen, wie z.B. Nomen, nachdem diese gesammelt wurden. Nomen können etwa in JAPE-Regeln als *NNP* abgefragt werden (siehe 4.x)
3. *Part-of-Speech-Tagging* - Erkennung von Satzbausteinen: Für ein genaues Verständnis, worum es in einer Anforderung geht, müssen die einzelnen Bestandteile des Textes zugeordnet werden. In unserem Programm werden gefundene Wörter etwa dahingehend geprüft, ob diese

ein System darstellen können, wenn sie zuvor als Nomen kategorisiert wurden.

Auf dieser Basis können dann passende Boilerplates identifiziert werden. Für die spätere Befüllung der Boilerplates mit den korrekten Begriffen ist die korrekte Erkennung von Worten im Satzzusammenhang daher wesentlich.

4. *Parsing* - Überführung in Boilerplates: Mithilfe des POS-Tagging wurden zuvor passende Boilerplates auf Basis gefundenen Annotationen innerhalb der Anforderung gefunden. Dies stellt nun die Grundlage für die Zusammensetzung der Boilerplates des entsprechenden Typs dar, indem Boilerplate-Objekte im Programm erzeugt werden.

Die Platzhalter werden dann von den Begriffen aus der Anforderung ersetzt, womit schlussendlich eine Anforderung in der formalen Sprache der Boilerplates vorliegt.

5. *Disambiguation* - Zuordnung der richtigen Konvertierung: Im Parsing werden zunächst alle möglichen Konvertierungen auf Basis möglicher Boilerplate-Typen durchgeführt. Nach der Erzeugung der Boilerplates liegen somit möglicherweise verschiedene Möglichkeiten der Deutung vor, von denen einzelne unpräzise oder fehlerhaft sein können.

Unpräzise Formulierungen können etwa durch eine hierarchische Anordnung der Boilerplates reduziert werden, wenn eine spezifische Boilerplate offensichtlich besser den Inhalt der Anforderung repräsentiert als eine eher allgemein formulierte. Die allgemeineren Boilerplates werden somit gleich ausgeschlossen, wenn eine bessere Konvertierung möglich ist. Siehe dazu auch Kap. 4.x.

Diese Unterschiede werden, sofern vorhanden, als mehrere Boilerplates im Programm dargestellt und an den Nutzer ausgegeben. Der Nutzer kann in der GUI dann die am besten zum ursprünglichen Anforderung passende Konvertierung auswählen.

6. *Context Module* - Bewertung im Zusammenhang: Bei der Identifikation der best passendsten Boilerplate wurden in Kap. 4.x Vorschläge zur Abbildung der Semantik gemacht. Diese Ansätze ermöglichen dann die Automatisierung der Auflösung bestimmter Mehrdeutigkeiten, wenn diese das Verständnis des Kontextes aus dem Zusammenhang mehrerer Sätze erfordern.

Im Kontext mehrerer Anforderungen kann der Nutzer im Rahmen des Review-Prozesses weitere Informationen über den Dokumentenkontext mit einfließen lassen. Dies kann etwa Informationen über vorige Anforderungen beinhalten, die bereits verarbeitet wurden, oder Meta-Informationen über das Anforderungsdokument wie etwa ergänzende, mitgeltende Unterlagen.

7. *Text Planning* - Speichern der korrekten Boilerplate: Nach der Auswahl von einer richtigen Konvertierung innerhalb der GUI hat sich der Nutzer für eine Boilerplate entschieden, die in das Export-Dokument übertragen werden soll. Durch diese Festlegung wird im Programm die entsprechende Boilerplate in die OutputList der Konvertierung eingetragen.

8. *Tactical Generation* - Erzeugung von Text aus BP-Objekten: Die Boilerplates liegen nach der Konvertierung zunächst als Objekte im Programm vor. Vor der Erzeugung des Ausgabe-Dokuments, das alle Anforderungen als Boilerplates enthält, müssen aus diesen Objekten der Text extrahiert werden.

Dieser Text liegt in den einzelnen Objekten als Zeichenkette vor und wird mithilfe einer Methode formatiert. Im Programm wird bei der Dokumentenerzeugung der Text aller übernommenen Boilerplates abgefragt.

9. *Morphological Generation* - Sprachliche Anpassung: In unserem ersten Entwurf der Prototypen wurde die Analyse auf englischen Text optimiert, da die Lastenhefte üblicherweise auf englisch verfasst sind (siehe Kap. 3.x). Die Boilerplates im Programm sind dementsprechend in englischer Sprache verfasst, wie auch das Ausgabe-Dokument.

Da der Text bereits innerhalb der Boilerplate-Objekte formatiert wurde bzw. vorliegt, entfällt hier eine aufwändige Anpassung an Sprache und ihre Grammatik.

10. *Output-Processing* - Erzeugung des Anforderungsdokuments aus Boilerplates: Nachdem die einzelnen Zeichenketten aus den Boilerplates extrahiert wurden, liegen nun die einzelnen Anforderungen in formaler Sprache als Text vor. Die Anforderungen müssen nun vor dem Export in der richtigen Reihenfolge angeordnet und an das Layout des Dokumentes angepasst werden.

Der Nutzer wählt dazu eine Datei über die GUI aus, in welche die Anforderungen exportiert werden. Der Dateityp ist dabei prinzipiell unbeschränkt, zur Vereinfachung wurde sich jedoch auf reine Textdateien (.txt) beim Export beschränkt.

Das Programm schreibt nun alle Anforderungen, getrennt durch Leerzeilen, in die Datei, sodass schlussendlich alle Anforderungen als Boilerplates exportiert sind.

Der Arbeitsablauf des Programms lässt sich anhand des Zustandswandlers in Abb. 4.x nachvollziehen. Dort ist die interne Verarbeitung unter Referenzierung der Stufen aus der NLP-Architektur als Prozess zwischen verschiedenen Akteuren dargestellt.

6.1 Auswertung der Testresultate

6.2 Mehrwert?

6.3 Ziel erreicht? Hypothese reviewen und schwafeln

Kapitel 7

Fazit

7.1 Zusammenfassung

7.2 Ausblick

Kapitel 8

Notizen

1. UML-Klassendiagramme im Vergleich zu Ontologien (Struktur, Beziehungen, Informatik-Bezug)
2. Nutzen von Ontologien, welcher Mehrwert kommt zur syntax durch sie hinzu?
3. Betriebliche Anforderungen an Tool: ACID aus DB?
4. Aktivitätsdiagramm aus PA mit GUI-User-Interaktion, Konkretisierung dortiger Activities
5. Use-Case Diagramm beim Konzept, PA nur referenzieren

Literaturverzeichnis

- [AWK06] Christian Allmann, Lydia Winkler, Thorsten Kölzow. „The requirements engineering gap in the OEM-supplier relationship.“, Journal of Universal Knowledge Management 1.2 (2006): 103-111.
- [CA15] Arora, Chetan, et al. „Automated checking of conformance to requirements templates using natural language processing.“ IEEE transactions on Software Engineering 41.10 (2015): 944-968.
- [BAL10] Helmut Balzert. „Lehrbuch der softwaretechnik: Basiskonzepte und requirements engineering.“, Springer-Verlag, 2010.
- [CAR52] Rudolf Carnap, „Meaning Postulates.“ , Philosophical Studies vol.3 S.65-73, 1952.
- [CHO57] Noam Chomsky, „Syntactic Structures“ , Mouton & Co., Feb. 1957.
- [COP04] Ann Copestake, University Of Cambridge, „Natural Language Processing“, 2004.
- [DGE19] Deutsche Gesellschaft für EMV-Technologie e.V. „CRS Customer Requirements Specification“, aus https://www.demvt.de/publish/view-full.cfm?ObjectID=ba9a0878_e081_515d_74a3411df6771be8, abgerufen am 09.01.19.
- [FF11] Farfeleder, Stefan, et al. „DODT: Increasing requirements formalism using domain ontologies for improved embedded systems development.“ Design and Diagnostics of Electronic Circuits & Systems (DDECS), 2011 IEEE 14th International Symposium on. IEEE, 2011.
- [FH14] Fockel, Markus, and Jörg Holtmann. „A requirements engineering methodology combining models and controlled natural language.“

- Model-Driven Requirements Engineering Workshop (MoDRE), 2014
IEEE 4th International. IEEE, 2014.
- [GM19] <https://gate.ac.uk/sale/tao/split.html> GATE User Guide, aufgerufen am 12.02.2019
- [HAO14] Hao Wu, et al. „ILLINOISCLOUDNLP: Text Analytics Services in the Cloud.“ LREC. 2014.
- [HE13] Alexander Heßeler, et al. „Anforderungsmanagement: Formale Prozesse, Praxiserfahrungen, Einführungsstrategien und Toolauswahl.“ Springer-Verlag, 2013.
- [HE18] HELLA Geschäftsbericht des Geschäftsjahres 2017/2018, von https://www.hella.com/hella-com/assets/media_global/2018.08.10_HELLA_Geschaeftsbericht_2018_geschuetzt.pdf, aufgerufen am 16.01.2019
- [HE19] Firma HELLA. „Unternehmensinformationen in Kürze.“, aus <https://www.hella.com/hella-com/de/HELLA-im-Ueberblick-723.html>, abgerufen am 09.01.19.
- [HL01] Hubert F. Hofmann, Franz Lehner. „Requirements engineering as a success factor in software projects.“ IEEE software 4 (2001): 58-66.
- [HP12] Andrea Harant, Miriam Poethen. „HOOD Capability Model für Requirements Definition und Requirements Management “ HOOD Ltd. 2012.
- [IR09] BGH, „Urteil vom 6. Juli 2000, I ZR 244/97“ Artikel beim Institut für Rechtsinformatik von der Universität des Saarlandes, 13. Oktober 2009
- [KN85] Kapur, Deepak, and Paliath Narendran. „An equational approach to theorem proving in first-order predicate calculus.“ ACM SIGSOFT Software Engineering Notes 10.4 (1985): 63-66
- [KR93] Ryan, Kevin. „The role of natural language in requirements engineering.“ Proceedings of the IEEE International Symposium on Requirements Engineering. IEEE, 1993.

- [LV10] F. Langenscheidt, B. Venohr. „Lexikon der deutschen Weltmarktführer: Die Königsklasse deutscher Unternehmen in Wort und Bild.“, Deutsche Standards–Gabal Verlag Google Scholar (2010).
- [MW02] Matthias Weber, Joachim Weisbrod. „Requirements engineering in automotive development-experiences and challenges.“, Requirements Engineering, 2002. Proceedings. IEEE Joint International Conference on. IEEE, 2002.
- [PR15] Pohl, Klaus, Chris Rupp. „Basiswissen Requirements Engineering: Aus-und Weiterbildung nach IREB-Standard zum Certified Professional for Requirements Engineering Foundation Level.“ dpunkt. verlag, 2015.
- [RS18] Felix Ritter, Aaron Schul. „Analyse aktueller NLP-Methoden und -Werkzeuge am Beispiel von GATE.“ Projektarbeit Fachhochschule Dortmund 2018, 18.11.2018
- [SB88] Muggleton, Stephen, Wray Buntine. „Machine invention of first-order predicates by inverting resolution.“ Machine Learning Proceedings 1988. 1988. 339-352
- [SHE07] L. Shen, G. Satta, A. K. Joshi. In Meeting of the Association for Computational Linguistics (ACL), „Guided learning for bidirectional sequence classification“ , 2007.
- [TOL09] Dhaval Thakker, Taha Osman, Phil Lakin. „Gate jape grammar tutorial.“ Nottingham Trent University, UK, Phil Lakin, UK, Version 1 (2009).
- [WE13] Weller, Katrin. „Ontologien.“ (2013): 207-218.
- [ZG02] Zowghi, Didar, and Vincenzo Gervasi. „The Three Cs of requirements: consistency, completeness, and correctness.“ International Workshop on Requirements Engineering: Foundations for Software Quality, Essen, Germany: Essener Informatik Beitiage. 2002.
- [ZH17] Zichler, Konstantin, Steffen Helke. „Ontologiebasierte Abhängigkeitsanalyse im Projektlastenheft.“ Automotive-Safety & Security 2017-Sicherheit und Zuverlässigkeit für automobile Informationstechnik (2017).