Freie Universität Berlin

Masterarbeit am Institut für Informatik der Freien Universität Berlin

Arbeitsgruppe Software Engineering

Qualität von Conclusions-Abschnitten: Eine qualitative Untersuchung

Ailis Oßwald
Matrikelnummer: 4997184
ailis@zedat.fu-berlin.de

Betreuer/in:	Prof. Dr. Lutz Prechelt
Eingereicht bei:	Prof. Dr. Lutz Prechelt
Zweitgutachter/in:	Prof. Dr. Claudia Müller-Birn

Berlin, 31. Oktober 2023

Zusammenfassung

Motivation: Conclusions-Abschnitte sind die letzten Abschnitte wissenschaftlicher Artikel, die die wichtigsten Schlussfolgerungen präsentieren. Sie enthalten zusätzlich einen studienbeschreibenden Einleitungsteil und einen Ausblick. Die Qualität von Conclusions-Abschnitten wissenschaftlicher Software-Engineering-Artikel nimmt ab. Es sind immer mehr Artikel vorhanden, in denen Autoren keine Schlussfolgerungen formulieren. Ohne Schlussfolgerungen wird kein Wissen präsentiert, sondern nur die Vorstufe dazu, was den Zweck der Wissenschaft verfehlt, Wissen zu teilen.

Ziel ist es zu verstehen, welche Elemente es innerhalb von Conclusions-Abschnitten gibt, die diese positiv bzw. negativ beeinflussen. Damit soll ein Verständnis über die Qualität von Conclusions-Abschnitten erlangt werden, aus dem Hinweise für das Verfassen hochqualitativer Conclusions-Abschnitte abgeleitet werden können.

Methoden: Dafür habe ich verschiedene Methoden angewandt, z.B. Inhaltsanalyse und offenes Codieren. Es sollte ein Codebuch mit objektiven Codes entstehen, mit welchem, die Qualität der Conclusions-Abschnitte beurteilt werden kann.

Ergebnisse: Insgesamt habe ich 48 Conclusions-Abschnitte analysiert und dabei ein objektives Codebuch erstellt. Elf Studierende haben dieses Codebuch getestet, indem sie quantitative Daten über Conclusions-Abschnitte gesammelt und analysiert haben. Aus mangelnden Rückschlüssen auf die Qualität, habe ich begonnen Qualitätsfaktoren auf Basis subjektiver Einschätzungen zu erarbeiten und fünf Faktoren gefunden. Einer ist bspw. die *Komplexität von Sätzen*.

Fazit: Es lassen sich mit strukturellen Inhalt-Codes kaum Rückschlüsse auf die Qualität von Conclusions-Abschnitten ziehen. Dies konnten die Studierenden mit der Analyse ihrer quantitativen Daten zeigen. Die gefundenen Qualitätskriterien fokussieren sich alle auf die enthaltenen Schlussfolgerungen der Conclusions-Abschnitte. Empfehlungen aus einer qualitative Forschung nur basierend auf veröffentlichten Conclusions-Abschnitten sind nicht ausreichend, um konstruktive Hinweise für hochqualitative Conclusions-Abschnitte abzuleiten.

Eidesstattliche Erklärung

Ich versichere hiermit an Eides Statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben, Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

31. Oktober 2023

Ailis Oßwald

Inhaltsverzeichnis

1 Einführung **7**
1.1 Anmerkungen zur Arbeit 9

2 Literaturrecherchen: Erwartungen an Conclusions-Abschnitte **9**
2.1 Literaturrecherche innerhalb der Wissenschaftstheorie 10
2.2 Literaturrecherchen im Softwareegineering 13
2.3 Erkenntnisse aus den Literaturrecherchen 16
2.3.1 Argumente für Schlussfolgerungen in Conclusions-Abschnitten 18

3 Entwicklung eines Codebuchs **20**
3.1 Methoden 20
3.1.1 Angewandte Techniken 20
3.1.2 Sonstiges 22
3.2 Datenbasis 23
3.3 Forschungsarbeitsbericht zur Entwicklung eines Codebuchs 23
3.3.1 Lernphase 24
3.3.2 Suche nach Arten von Conclusions-Abschnitten 29
3.4 Ergebnisse: Das Codebuch 31

4 Erprobung des Codebuchs durch Studierende **36**
4.1 Forschungsarbeitsbericht über das Studierendenprojekt 37
4.1.1 Vorarbeiten zum Kurs 37
4.1.2 Arbeiten während des Kurses 39
4.1.3 Nacharbeiten zum Studierendenprojekt 41
4.2 Ergebnisse des Studierendenprojekts – quantitative Analyse 43
4.2.1 Beurteilung der Ergebnisse der Studierenden 47
4.2.2 Reflexion zum Studierendenprojekt 48
4.3 Erkenntnisse aus dem Studierndenprojekt 51
4.3.1 Erkenntnisse zum Codebuch 51
4.3.2 Aufbau von Conclusions-Abschnitten 53

5 Suche nach Qualitätsfaktoren **55**
5.1 Vergleichen von Extremen 56
5.2 Überprüfung meiner gefundenen Qualitätsfaktoren 59
5.2.1 Ergebnisse der Studie 61
5.3 Erkenntnisse: Was beeinflusst die Qualität von Conclusions-Abschnitten? 65
5.3.1 Faktoren die komplette Conclusions-Abschnitte betreffen 70
5.3.2 Faktoren die Schlussfolgerungs-Sätze betreffen 71
5.3.3 Wechselwirkungen zwischen den Faktoren 75
5.3.4 Einschränkungen der Erkenntnisse 77

6 Fazit **78**
6.1 Selbstreflexion 80
6.2 Ausblick 82

A Codes zu Beginn der offenen Codierungsphase **91**

B Aufgabenstellungen für die offene Codierungsphase der Studierenden: **91**

C Dokumente zur informellen Studie zur Überprüfung der gefundenen Qualitäts-Faktoren **92**

C.1 Genutzte Beispiele für ganze Conclusions-Abschnitte 92
C.2 Genutzte Beispiele für einzelne Schlussfolgerungssätze 93
C.3 Vorbereitetes Dokument zum festhalten von Notizen 95
C.4 detaillierte Ergebnisse 98

1 Einführung

Auf dem Gebiet des Software-Engineering (SWE) werden jährlich mehrere tausend Artikel in Konferenzen veröffentlicht. 2012 waren dies grob 4.000 Artikel, Tendenz steigend [29]. Dadurch scheint es naheliegend, dass jede Person einen einfachen Zugang zu einer immensen und weiter wachsenden Menge an SWE-Wissen haben müsste. Einfacher Zugang bedeutet hier, dass eine Person, die Zugriff auf einen Artikel im Volltext hat, nach dem simplen Akt des Lesens des Artikels um das Wissen der Studie schlauer ist.

In der Realität scheint dies nicht der Fall zu sein. Nach dem Lesen des Artikels wissen die Personen meist viel über den Studienaufbau und die -durchführung, nicht aber was die konkreten Erkenntnisse sind. Das bedeutet, Ziele, Methoden und Beobachtungen werden genannt, nicht aber die Schlussfolgerungen der Autoren. Nicht alle Artikel vermitteln also neues Wissen in Form von Erkenntnissen, sondern nur Informationen, die aufmerksame und mutige Lesende mit etwas Spekulation selbst zu Schlussfolgerungen und damit zu Erkenntnissen verarbeiten müssten. Da kaum Lesenden die Studie selber durchgeführt haben, stehen ihnen in jedem Fall weniger Informationen zur Verfügung als den Autoren. Lesende können daher nur Schlussfolgerungen auf weniger genauen Ergebnissen tätigen.

Obwohl immer mehr Forschung gemacht und veröffentlicht wird, wird häufig vermieden zu Schlussfolgern und damit die Gelegenheit verpasst, konkretes Wissen zu teilen.

Dass tendenziell weniger SWE-Artikel schlussfolgern, ist Prof. Dr. Lutz Prechelt aufgefallen, der an der Freien Universität Berlin doziert und arbeitet. Als Professor und Forscher im Bereich SWE liest er regelmäßig wissenschaftliche Publikationen, die ihn in letzter Zeit zunehmend mit einem unbefriedigenden Gefühl zurückließen. Mit anderen Worten, Prof. Dr. Prechelt schien die Qualität der SWE-Artikel abgenommen zu haben und teilweise nicht mehr ausreichend zu sein. Einen Grund dafür vermutet er im Abstract, dem ersten und am stärksten standardisierten Teil eines Artikels. Ein weiterer Grund sei die mangelnde Motivation der Autoren, ausformulierte Schlussfolgerungen zu veröffentlichen.

Verwandte Studie „qabstracts“ Um seine Vermutungen zu überprüfen und zu belegen, startete Prof. Dr. Prechelt eine Inhaltsanalyse über Abstracts von SWE-Artikeln (qabstracts-Studie). In dieser Forschungsarbeit wird die Methode des offenen Codierens verwendet. Es wird ein Codebuch entwickelt, mit dem quantitative Daten über die Struktur von Abstracts gesammelt werden, die anschließend einen Rückschluss auf die Qualität der Abstracts zulassen. Ein Beispiel für einen solchen Code, mit Hilfe dessen sich die Qualität einschätzen lässt, ist das „Understandability gap“. Mit diesem Code werden Stellen gezählt, bei denen das Abstract mehrdeutig oder unverständlich ist. Je mehr solcher Stellen im Abstract enthalten sind, desto schlechter ist die Qualität des Abstracts. Das Abstract ist der allererste Teil des Artikels den Interessierte lesen. Deswegen sollte das Abstract verständlich sein und keine „Understandability gaps“ enthalten.

Am Ende der Studie sollte es möglich sein, eine Aussage über die allgemeine Qualität von Abstracts in SWE-Artikeln zu machen. Mit den dann voraussichtlich quanti-

tativ belegten Aussagen über die nach Prof. Dr. Prechelts Vermutung weit verbreitete schlechte Qualität von Abstracts, sollen die Forschenden des SWE auf die schlechte Qualität aufmerksam gemacht werden und zu besseren Abstracts angeregt werden.

Die qabstracts-Studie ist noch nicht veröffentlicht. Der aktuelle Stand der Forschung dazu ist unter folgendem Link in einem privaten Repository zu finden: https://github.com/serqco/qabstracts/tree/main.

Das Repository bleibt bis zur Veröffentlichung der Studie privat. Interessierte können sich an Prof. Dr. Prechelt wenden (prechelt@inf.fu-berlin.de).

Ziel Es soll eine ähnliche Studie, wie die qabstracts-Studie über Schlussfolgerungen in wissenschaftlichen Artikeln durchgeführt werden. Diese Folgestudie soll dazu beitragen, dass Forschende des SWE dazu angeregt werden qualitativ hochwertigere Artikel zu schreiben. Eine grundlegende Annahme dieser Studie ist, dass Conclusions-Abschnitte die wichtigsten Schlussfolgerungen enthalten sollten. Die vorliegende Masterarbeit ist der erste Teil dieser Studie. Das erste Ziel war es, die Qualität von Conclusions-Abschnitten beurteilbar und quantifizierbar zu machen. Im Idealfall sollten dabei alle Abschnitte eines Artikels beachtet werden, die potentiell Schlussfolgerungen enthalten könnten. Da dies den Aufwand einer Masterarbeit sprengen würde, beschränkt sich die vorliegende Arbeit auf den Conclusions-Abschnitt, der die Aufgabe hat, die Erkenntnisse der Studie zusammenzufassen und dem Leser die wichtigsten Schlussfolgerungen aus der Studie zu präsentieren.

Am Ende soll ein fertiges Codebuch für Conclusions-Abschnitte stehen, mit dem die Qualität solcher Abschnitte hinreichend bewertet werden kann. Zusätzlich soll ein Ansatz herausgearbeitet haben, wie Schlussfolgerungen im gesamten Artikel systematisch gefunden werden können.

Hauptziel der Arbeit ist ein besseres Verständnis der Qualität von Conclusions-Abschnitten zu vermitteln.

Die Qualität von Conclusions-Abschnitten im SWE ist ein kaum behandeltes Thema (2.2). Daher werde ich in diesem Dokument versuchen, nicht nur die Methoden und Erkenntnisse zur Qualität von Conclusions-Abschnitten zu beschreiben, sondern auch alles weitere, was nachfolgenden Forschenden helfen könnte dieses Gebiet weiter zu erforschen. Dazu gehören Berichte über erfolgreiche und erfolglose Methoden sowie Beschreibungen unausgereifter und falscher Hypothesen. Da es sich um eine Masterarbeit handelt, werde ich auch persönliche Schwierigkeiten und Erkenntnisse einfließen lassen. Dies könnte z.B. für Personen hilfreich sein, die zum ersten Mal mit qualitativer Forschung in Anlehnung an die „Grounded Theory Methodology“ arbeiten, wie ich es tat.

Gliederung der Arbeit Chronologisch lief die Arbeit wie folgt ab: Ich begann mit dem Erlernen von Techniken der „Grounded Theory Methodology“ (3.1), dabei wurde die inhaltliche Struktur von Conclusions-Abschnitten fokussiert (3). Im Anschluss habe ich mit einer Gruppe von Studierenden am Codebuch zusammengearbeitet (4). Nachfolgend habe ich die Unterschiede zwischen Extremen untersucht (5) und mit einer Literaturrecherche (2) und dem Verfassen dieser Arbeit abgeschlossen.

Bis auf die Literaturrecherche folgt dieses Dokument der chronologischen Zeitabfolge. Alle Forschungsarbeiten zum Thema dieser Arbeit sind im jeweiligen Abschnitt

beschrieben und so strukturiert, dass ggf. mit einer Beschreibung von Methoden begonnen, mit einem Forschungsbericht inkl. Ergebnissen fortgefahren wird und zuletzt mit einer Zusammenfassung abgeschlossen wird, die Erkenntnisse zum Thema und eine Reflexion zur Arbeitsweise enthält.

Ich beginne mit einer Literaturrecherche zu Conclusions-Abschnitten (2). Dabei suche ich innerhalb der Wissenschaftstheorie und des SWE nach Quellen, um Begründungen oder Gegenargumente für Schlussfolgerungen innerhalb von Conclusions-Abschnitten zu finden (2.3.1). Im Anschluss wird der erste Teil meiner Forschung beschrieben (3), in dem ich mich allgemein orientieren musste und mit dem Erlernen von offenem Codieren (3.3), dem Erstellen eines Codebuchs (3.4) und dem Verstehen von Conclusions-Abschnitten beschäftigt war. Das Codebuch wurde im Anschluss von einer Studierendengruppe weiterentwickelt und erprobt, indem auf Basis des Codebuchs quantitative Daten gesammelt wurden (4). Die Ergebnisse aus der Quantifizierung (4.2) lassen Aussagen über die Struktur von Conclusions-Abschnitten zu (4.3.2). Die Erprobung des Codebuchs ergab außerdem, dass die Struktur von Conclusions-Abschnitten nicht genug über die Qualität aussagt. Um Qualitätsfaktoren zu finden, habe ich daher die Technik des Vergleichens von subjektiv festgelegten Extremen angewandt (5.1) und meine gefundenen Faktoren mit einer informellen Studie getestet (5.2). Die erarbeiteten Qualitätsfaktoren und Erkenntnisse der Studie sind in Unterabschnitt 5.3 zu finden. Im letzten Teil der Arbeit ziehe ich ein Fazit (6) mit Bezug auf die Zielstellung und über meine erbrachte Arbeit und beende das Dokument mit einem Ausblick.

1.1 Anmerkungen zur Arbeit

In der vorliegenden Arbeit werde ich Abkürzungen verwenden, die allgemein bekannt sind.

Da die meisten wissenschaftlichen Arbeiten in englischer Sprache publiziert und gelesen werden, ist die gesamte Datenbasis auf Englisch. Dadurch, dass diese Masterarbeit auf Deutsch verfasst ist, kann das zuweilen dazu führen, dass englische Begriffe verwendet werden, weil es kein ausreichend passendes deutsches Äquivalent gibt, z.B. für den Begriff „Conclusions-Abschnitt“. Eine (engl.) „Conclusion“ ist entweder eine Entscheidung oder ein Urteil über eine ausreichend untersuchte Sache oder der letzte Teil von etwas, wie z.B. ein Schluss [10]. Da diese Zweideutigkeit des Begriffes relevant ist, werde ich im Folgenden den englischen Begriff weiter nutzen.

2 Literaturrecherchen: Erwartungen an Conclusions-Abschnitte

Insgesamt habe ich drei Literaturrecherchen durchgeführt. Eine im Bereich der Wissenschaftsphilosophie (2.1) und zwei im Feld des SWE (2.2). Ziel der ersten Literaturrecherche war es, den Zweck von Conclusions-Abschnitten und Schlussfolgerungen in der Wissenschaft zu verstehen (Unterabschnitt 2.3). Mit den verbleibenden Literaturrecherchen sollte ein allgemeines Bild von Erwartungen an Conclusions-Abschnitte im SWE erstellt werden (Abschnitt 2.2).

Eigentliches Ziel der Recherchen war es, die Annahme, auf der diese Forschung beruht, zu bestätigen. Diese Annahme lautet, dass „jede Publikation einen Conclusions-

Abschnitt haben sollte, in dem die wichtigsten Erkenntnisse präsentiert werden“. Leider ließen sich keine Quellen finden, die genau diese Aussage treffen.

2.1 Literaturrecherche innerhalb der Wissenschaftstheorie

Im Folgenden werde ich als Erstes die genutzten Suchparameter aufzählen und dann auf das allgemeine Konzept und den Zweck von Wissenschaft und Schlussfolgerungen und deren Zusammenhang qabstracts-Studie eingehen. Ich erkläre warum Schlussfolgerungen ein wesentlicher Bestandteil wissenschaftlicher Arbeiten sind und warum Sie in Conclusions-Abschnitten enthalten sein sollten. Antworten auf diese Fragen habe ich in der Wissenschaftstheorie gesucht. Dies erschien mir passend, da sich die Wissenschaftstheorie in normativer und deskriptiver Weise mit Wissenschaft beschäftigt [12, 49]. Das bedeutet, dass Verfahrensweisen und Vorgehensweisen von Forschern thematisiert werden [55]. Dazu gehört auch das Thema des Verfassens wissenschaftlicher Publikationen.

Methode Ich habe nach Artikeln gesucht, die konkrete Aussagen zu Conclusions-Abschnitten und deren Struktur machen oder den Aufbau wissenschaftlicher Publikationen besprechen. Dazu habe ich die folgenden Stichwörter und Kombinationen davon verwendet: „purpose articles“, „structure articles“, „conclusion section“, „how to write article“, „guide article“, „purpose“, „journal article“, „structure“, „normative article“. Das Wort „article“ habe ich variiert mit „paper“, „study“ und ergänzt mit „scientific“ oder „journal“.

Die Suche erfolgte über die jeweilige Stichwortsuche auf der Website der Konferenz. Folgende Konferenzen habe ich durchsucht: „Journal for General Philosophy of Science“ [41], „Science and Philosophy“ [71], „Philosophy of science“ [63], „The british journal for the philosophy of science“ [83], „European Journal for philosophy and science“ [26], „Erkenntnis“ [23], „International studies in the philosophy of science“ [39], „Metascience“ [58], „Perspectives on science“ [62] und „Synthese [82]“. Die Suche fand am 29.08.2023 und 30.08.2023 statt und umfasste alle online verfügbaren Artikel. Pro Suchergebnis habe ich die ersten 30 bis 60 Artikel nach Titel und Vorschau des Abstracts ausgewählt, was immer ca. drei Ergebnisseiten entsprach. Wenn auf der ersten Seite nicht annähernd passende Artikel gefunden wurden, habe ich die Suche bereits nach 15 bis 20 Treffern beendet und eine neue Suche mit anderen Suchbegriffen gestartet.

Ergebnisse - Schlussfolgerungen und Zweck der Wissenschaft: Nach der Wortbedeutung ist eine *Schlussfolgerung* im wissenschaftlichen Kontext die Bewertung eines ausreichend untersuchten Sachverhaltes oder die Formulierung einer Erkenntnis auf der Grundlage von Daten. Es gibt zwei Arten von Schlussfolgerungen: Schlussfolgerungen, die verallgemeinern, und Schlussfolgerungen, die einen Existenzbeweis liefern. Beide Arten von Schlussfolgerungen sind ausformulierte Erkenntnisse der Autoren, die auf Basis der Forschungsergebnisse formuliert werden. Solche ausformulierten Erkenntnisse sind die Essenz der Wissenschaft. In der Wissenschaft geht es um Wissen, und zwar um das Streben nach mehr Wissen im Hinblick auf ein besseres Verständnis unserer Welt [52, 69]. Wissen ist ein Produkt des Verstehens,

des Versuchs, System und Ordnung zu erkennen [52] und „rationale wahrheitsorientierte Überzeugungsbildung“ [78] zu betreiben. Wissenschaftliches Arbeiten ist die systematische und geordnete Suche nach Wissen [52], um zu begründetem Wissen zu gelangen [20, 49]. Jede Schlussfolgerung am Ende wissenschaftlicher Forschung ist begründetes Wissen und damit immer mindestens ein Existenzbeweis und Anstoß für neue Forschungen.

Die Wissenschaftstheorie besagt, dass Wissen durch Hypothesen gewonnen wird, die bestätigt, widerlegt oder angepasst werden [49, 12, 20, 85, 50]. Bestätigte bzw. nicht widerlegte Hypothesen sind Theorien. Hypothesen werden z.B. aufgestellt, um Beobachtungen zu verallgemeinern und damit allgemeine Phänomene und Gesetzmäßigkeiten beschreiben zu können [52, 1, 9]. Solche Verallgemeinerungen sind Schlussfolgerungen und das bedeutet, dass Schlussfolgerungen Wissen sind.

Zusammengefasst besteht das Ziel der Wissenschaft darin, neue Erkenntnisse zu gewinnen, die unsere Welt etwas besser beschreiben als bisher. Erkenntnisse sind neues Wissen in Form von ausreichend durch Forschungsergebnisse begründeten Theorien. Schlussfolgerungen in der Wissenschaft sind solche Erkenntnisse. Sie können die Form von Verallgemeinerungen annehmen. Schlussfolgerungen zu ziehen und zu präsentieren sollte das Ziel jeder Forschung sein.

Ergebnisse - Begründungen für Conclusions-Abschnitte mit Schlussfolgerungen: Eine wichtige Annahme meiner Forschung ist, dass Conclusions-Abschnitte ein essentieller Teil von wissenschaftlichen Artikeln sein sollten, da in diesen Abschnitten die wichtigsten Ergebnisse des Artikels präsentiert werden sollten.

Innerhalb der Wissenschaftstheorie ließen sich keine Artikel finden, die zuvor genannte Annahme konkret bestätigten oder sich konkret mit der Struktur oder dem Verfassen eines wissenschaftlichen Artikels befassen. Möglicherweise liegt das an fehlenden Fachwissen, dass mir bestimmte Werke und möglicherweise auch Fachbegriffe nicht bekannt sind.

Ich habe jedoch Artikel gefunden, die sich mit grundlegenderen Fragen der Wissenschaftstheorie befassen und damit Hinweise auf die Struktur wissenschaftlicher Artikel enthalten sollten und damit entweder für oder gegen die Existenz von Conclusions-Abschnitten argumentieren.

Die folgende Liste fasst Begründungen für Schlussfolgerungen in Conclusions-Abschnitten zusammen.

- **Schlussfolgerungen sind keine Resultate:** Resultate sind keine Erkenntnisse [14]. Resultate sind beobachtete Fakten, die nur Aussagekraft über genau die Situation haben, in der sie beobachtet wurden. Resultate sind Einzelfälle, die ein Baustein zur Erklärung von Phänomenen sein können, aber selber nie diese Erklärung sind.

  Warum genügt es nicht einfach, die Resultate mitzuteilen und jedem Lesenden selbst, die Schlussfolgerungen zu überlassen? Es ist nicht möglich, alle Details einer Studie in einem Artikel vollständig wiederzugeben. Die Forschenden, die die Studie durchgeführt haben, besitzen das umfangreichste Wissen zu ihrer Studie und sind daher am besten in der Lage, Schlussfolgerungen zu ziehen

[8]. Selbst wenn dies nicht der Fall wäre, wäre es ungünstig den Lesenden das Schlussfolgern zu überlassen. Denn einen Artikel genügend zu verstehen und nachzuvollziehen um selber Schlussfolgerungen zu machen, ist in jedem Fall aufwändiger, als Schlussfolgerungen präsentiert zu bekommen. Um zu Schlussfolgerungen zu gelangen, ist ein Gedankenprozess notwendig, der eine gewisse Zeit in Anspruch nimmt [8]. Zu Beginn steht der präsentierte Inhalt, danach erfolgt eine Transition und am Ende steht der resultierende Inhalt [8]. Beispiele für Transitionen sind z. B. Analyse, Klassifizierung, Sortierung, Nummerierung, Messung, Korrelation usw [8]. Der Inhalt muss bereits vorhanden sein und darf nicht erst neu gewonnen werden, denn dies würde genau den gleichen Prozess wiederholen und damit den Denkprozess unweigerlich verlangsamen. Wenn alle Lesenden bei jedem Artikel den gesamten Denkprozess von der Präsentation der Beobachtungen bis zu den Schlussfolgerungen selbst durchführen müssten, wäre die Grundlage für die Entwicklung neuer Theorien wahrscheinlich ungemein langsam. Natürlich meine ich damit nicht, dass Erkenntnisse bedingungslos als wahr und richtig angesehen werden sollten. Es ist jedoch weniger aufwendig, wenn man überprüft, ob die mitgelieferten Argumente valide sind und dann entsprechend die Erkenntnisse nutzt, als diese Erkenntnisse selbst zu formulieren.

- **Lesbarkeit des Artikels:** Wie bereits erwähnt, ist es das Ziel der Wissenschaft zu mehr Wissen zu gelangen. Dazu gehört das Verbreiten des Wissens. Wissen ist nur dann relevant, wenn es weiter verteilt wird. Dies geht aus der Idee des kritischen Rationalismus von Popper [64] hervor und lässt sich in anderen Werken ebenfalls auffinden [50, 35]. Eine relevante Form der Informationsverbreitung in der Wissenschaft sind Artikel. Ziel solcher Artikel ist es, dem Leser die neu gewonnenen Erkenntnisse überzeugend, d.h. hinreichend begründet, zu vermitteln [70, 81]. Artikel enthalten daher die beste Argumentation für Gültigkeit und Relevanz der Forschung [70, 81]. Zu diesem Zweck enthält beispielsweise jeder Artikel einen Abschnitt über die genutzten Methoden, in dem die systematische Vorgehensweise des Forschers begründet und erläutert wird [35]. Da das Hauptziel von Artikeln das Verbreiten von Wissen ist, sollte es eigene Abschnitte geben deren Hauptzweck das weitergeben von Wissen ist. Einer dieser wissensdedizierten Abschnitte scheint der Diskussion-Abschnitt zu sein, der ausführlich ist und dessen wichtigste Schlussfolgerungen in einem Conclusions-Abschnitt zusammengefasst werden sollten. Mit einem Conclusions-Abschnitt hätten Autoren die Möglichkeit, das Wichtigste zu wiederholen und losgelöst von den Details der Forschung darzustellen. Die Chance, dass jeder Leser das Wichtigste aus dem Artikel mitnimmt, würde stark erhöht und damit die Verbreitung des intendierten Wissens durch den Artikel deutlich verbessert werden.

- **Am Ende auf den Anfang zurückkommen:** Ein wissenschaftlicher Artikel präsentiert Erkenntnisse bezüglich eines zu Beginn festgelegten Themas und liefert im Laufe des Artikels Argumente für diese Erkenntnisse. Damit ähnelt ein wissenschaftlicher Artikel im Aufbau entfernt einer argumentativen Rede, die wie der Artikel selbst eine eingangs gestellte Frage begründet zu beantworten versucht. Rhetorikern nach sollten Reden typischerweise immer mit einer

Bezugnahme des Redners auf das eingangs dargestellte Problem enden [35]. Bezogen auf wissenschaftliche Artikel, würde das heißen, dass der letzte Abschnitt eines Artikels immer die gewonnenen Erkenntnisse und damit die wichtigsten Schlussfolgerungen der Forschung enthalten sollte, da diese immer die Forschungsfrage beantworten.

- **Effiziente Wissensverbreitung:** Seit dem 20. Jahrhundert ist die Struktur wissenschaftlicher Artikel sehr einheitlich [35]. Dies spiegelt wider, dass Artikel heutzutage eher als Ressource genutzt werden, in der Informationen gesucht werden, als dass sie als Zeitschriftenartikel gelesen werden [35]. Artikel sind darauf ausgelegt, Informationen möglichst effektiv zu vermitteln. Es wäre daher naheliegend, dass jeder Artikel den gleichen Abschnitt verwenden würde, um seine Hauptaufgabe, die Wissensverbreitung, zu erfüllen. Ein passender Abschnitt dafür wäre, allein dem Namen nach, der Conclusions-Abschnitt.

2.2 Literaturrecherchen im Softwareegineering

Ziel dieser beiden Literaturrecherchen war es konkret, SWE-Artikel zu finden, in denen die Struktur von wissenschaftlichen Artikeln besprochen wird, inklusive Conclusions-Abschnitten. Dafür habe ich zu Beginn der Arbeit gesucht und separat davon mit neuen Ideen zum Ende der Arbeit. Die Literaturrecherchen werden chronologisch abgehandelt. Dabei wird jeweils zuerst die Umsetzung beschrieben und anschließend die Ergebnisse.

1. Suche zu Conclusions-Abschnitte im SWE Ein weiteres Ziel der ersten Recherche war es, mich mit dem Thema vertraut zu machen, mehr über Schlussfolgerungen zu erfahren und eine grobe Vorstellung davon zu bekommen, was im SWE von Conclusions-Abschnitten erwartet wird. Diese Suche nach Artikeln erfolgte über Google Scholar und mit folgenden Stichwörten und Kombinationen aus diesen: „how to read“, „articles“, „guide write“, „how to write“, „Softwareengineering“, „conclusion section“. Die Suche mit „how to read articles“ war erfolgreich. Suchen mit „write“ bzw. „guide“ und „Softwareengineering“ haben viele unpassende Ergebnisse erzeugt, da im SWE nicht nur Artikel, sondern auch Code geschrieben wird. Für das Schreiben von Code gibt es viele Quellen, die Empfehlungen, Praktiken und Methoden beschreiben. Die Suche mit dem Stichwort „Conclusion section“ ergab keine passenden Artikel. Suchen ohne „Softwareengineering“ haben passendere Artikel ergeben, die sich konkret mit dem Lesen bzw. Schreiben von wissenschaftlichen Publikationen befasst haben, allerdings kamen diese hauptsächlich aus dem Bereich der Medizin. Auf passende Artikel habe ich jeweils Vorwärts- und Rückwärtssuche angewandt und konnte somit die Menge relevanter Artikel ungefähr verdoppeln. Den Inhalt der Artikel habe ich teils überflogen und teils konkret nach dem Stichwort „conclusion“ und Varianten davon durchsucht. Gelesen habe ich hauptsächlich Textstellen, die die Stichwörter enthielten.

Ergebnisse der ersten SWE-Literaturrecherche: Die Artikel, die ich bei dieser ersten Suche gefunden habe, befassen sich zum Teil mit der Anwendung und schriftlichen

Darstellung von Forschungsmethoden und zum Teil eher allgemein damit, wie wissenschaftliche Artikel publiziert werden. Nicht alle Artikel stammen aus dem Bereich des SWE. Die Artikel, die sich konkret mit dem Schreiben oder Lesen eines wissenschaftlichen Artikels befassen, stammen überwiegend aus der Medizin.

Passende Artikel, die aus dem SWE stammen, besprechen Forschungsmethoden und die Art und Weise wie Schlussfolgerungen durch Daten unterstützt und gefunden werden können [68, 72, 73, 45, 88, 21, 47, 59]. Dabei wird impliziert, dass am Ende jeder Studie mindestens eine Schlussfolgerung niedergeschrieben werden sollte. Es wird aber nicht zwingend spezifiziert, in welchem Abschnitt das am passendsten wäre. Ein weiteres Ergebnis ist, dass keiner der Artikel angedeutet hat, dass Schlussfolgerungen nicht nötig seien. Von den fünf gefundenen SWE-Artikeln, die Näheres zu Conclusions-Abschnitten beschreiben, erwarten alle in einem solchen Abschnitt Schlussfolgerungen [47, 21, 73, 40, 59]. Von diesen Artikeln erwarten zwei Schlussfolgerungen ausschließlich im Conclusions-Abschnitt [47, 21], zwei andere Artikel erwarten, dass es neben dem Conclusions-Abschnitt mit Schlussfolgerungen noch weitere Abschnitte gibt, die Schlussfolgerungen ausführlicher enthalten [73, 40]. Ein Artikel hält einen Diskussion-Abschnitt für Pflicht und einen Conclusions-Abschnitt für optional [59].

Von den gefundenen Artikeln aus Medizin und Naturwissenschaften, die erklären, wie schriftliche Publikationen zu schreiben oder zu lesen sind, erwarten alle sechs, dass Schlussfolgerungen enthalten sind [22, 80, 42, 54, 57, 87]. Drei der Artikel erwarten, dass es einen Conclusions-Abschnitt zusätzlich zu einem Diskussion-Abschnitt gibt [87, 54, 22]. Der Unterschied von Diskussion-Abschnitten zu Conclusions-Abschnitten ist, dass letztere nur die wichtigsten Schlussfolgerungen enthalten. Zwei der medizinischen Artikel erwarten einen Diskussion-Abschnitt und sehen den Conclusions-Abschnitt als optional an [80, 42]. Ein medizinischer Artikel erwartet lediglich einen Diskussion-Abschnitt [57].

Die Literaturrecherche hat gezeigt, dass Conclusions-Abschnitte nicht immer selbstverständlich sind und oft gar nicht erwähnt werden. Wenn sie jedoch inhaltlich beschrieben werden, wird immer erwartet, dass sie Schlussfolgerungen enthalten. Insgesamt implizieren zumindest alle Artikel, dass es Schlussfolgerungen geben muss. Die meisten Artikel erwarten eine detaillierte Beschreibung der Schlussfolgerungen in einem Diskussion-Abschnitt. Ein Artikel beschreibt verschiedene Formen wissenschaftlicher Artikel und weist darauf hin, dass die Struktur der Artikel und damit die Verortung der Schlussfolgerungen stark von der Zeitschrift abhängt, in der er publiziert wird [18].

2. Literaturrecherche zu Conclusions-Abschnitten im SWE: Gesucht habe ich diesmal ausschließlich innerhalb von SWE-Konferenzen: „Empirical Software Engineering“, „Information and Software Technology“ und allen Zeitschriften, die in der „ACM“- und „IEEE“-Bibliothek enthalten sind. Das bedeutet, dass alle Artikel aus SWE nahen Bereichen der Wissenschaft stammen. Dabei habe ich die integrierten Suchfunktionen der Webseiten der Konferenzen genutzt und mit folgenden Stichwörtern gesucht: „how to read“, „write scientific“, „guide write“, „structure“ - jeweils in Kombination mit mindestens einem der Begriffe „paper“, „article“ oder „scientific“. Ausgewählt habe ich hier nach Überschriften und im Anschluss nach Abstracts.

Innerhalb der Artikel habe ich Paragraphen, die einzelne oder mehrere Schlüsselwörter beinhalten, näher untersucht: „conclusion“, „learning“, „take away“, „lesson“, „implication“. Nach einer inhaltlichen Durchsicht erschien nur ca. ein Drittel der ausgewählten Artikel passend. Auf die verbliebenen passenden Artikel habe ich Vor- und Rückwärtssuche angewandt.

Ergebnisse der zweiten SWE-Literaturrecherche: Alle gefundenen SWE-Artikel konstatieren eindeutig, dass Schlussfolgerungen ein obligatorischer Bestandteil wissenschaftlicher Artikel sind. In drei Artikeln werden Schlussfolgerungen konkret als Verallgemeinerungen beschrieben, die notwendig sind, um aus Resultaten Wissen zu kreieren [40, 73, 46].

Schlussfolgerungen sollten so formuliert sein, dass Lesende, kein eigenes Wissen anwenden müssen um die Schlussfolgerungen zu verstehen. Das bedeutet, Implikationen müssen ausformuliert werden und Beobachtungen müssen genau beschrieben werden [67]. Schlussfolgerungen sollten direkt und einfach formuliert werden [67].

Vier Artikel erwarten Schlussfolgerungen im Detail im Diskussion-Abschnitt und in zusammengefasster Form auch im Conclusions-Abschnitt [67, 7, 51, 61]. Zwei ältere Artikel aus den Jahren 1960 und 1977 erwarten Schlussfolgerungen nur innerhalb des Diskussion-Abschnitts, aber gehen auch nicht von der Existenz eines Conclusions-Abschnitts aus [17, 38]. Zwei der gefundenen Artikel erwarten Schlussfolgerungen ausschließlich in einem Conclusions-Abschnitt, erwähnen aber auch keinen separaten Diskussion-Abschnitt [53, 77]. Die Conclusions-Abschnitte in diesen Artikeln gleichen inhaltlich stark den Diskussion-Abschnitten heutiger Artikel [17, 38]. Bei einem Artikel wird beschrieben, dass ein Conclusions-Abschnitt mit Schlussfolgerungen erwartet wird, nicht aber, ob dieser einen Diskussion-Abschnitt ersetzt oder ergänzt [43].

Alle Artikel die einen Conclusions-Abschnitt näher beschreiben, erwarten dass dieser Schlussfolgerungen enthält [67, 77, 53, 7, 43].

Auch die zweite Literaturrecherche zeigt, dass Schlussfolgerungen ein selbstverständlicher Bestandteil wissenschaftlicher Forschung und Publikationen zu sein scheinen. Ein Abschnitt mit Schlussfolgerungen sollte immer im Artikel enthalten sein, und wenn es einen Conclusions-Abschnitt gibt, sollte dieser immer Schlussfolgerungen enthalten. Allerdings scheint es häufig der Fall zu sein, dass kein abschließender kürzerer Abschnitt erwartet wird, der die wichtigsten Schlussfolgerungen zusammenfasst, sondern dass ein ausführlicher Abschnitt mit Schlussfolgerungen erwartet wird. So ein ausführlicher Abschnitt mit Schlussfolgerungen entspricht dem heutigen Verständnis von Diskussion-Abschnitten und anscheinend einem älteren Verständnis von Conclusions-Abschnitten.

Fazit beider Literaturrecherchen: Die Literaturrecherchen bestätigen die wissenschaftstheoretische Aussage, dass das Ziel der Forschung Wissen ist und dass dieses Wissen verbreitet werden sollte. Alle Artikel erwarten Schlussfolgerungen. Ein Artikel betont darüber hinaus, dass die Veröffentlichung von Artikeln das Ziel jeder Forschung ist [17]. Weitere Artikel aus der Literaturrecherche, die sich nach genauer Prüfung als ungeeignet herausstellten, beschäftigen sich nur damit, wie man Artikel erfolgreich veröffentlicht und formuliert, um Wissen möglichst weit zu verbreiten [25, 44, 76].

Es gibt keine einheitlichen Aussagen über Conclusions-Abschnitte. Schlussfolgerungen werden von allen Artikeln erwartet, aber in verschiedenen Abschnitten: Die Mehrheit der Artikel (10/14) erwartet einen Abschnitt mit Schlussfolgerungen, der einem ausführlichen Diskussion-Abschnitt entspricht. Ein Conclusions-Abschnitt, der nur die wichtigsten Schlussfolgerungen enthält, wird in (6/14) der Artikel vorausgesetzt und in einem Artikeln als optional angesehen.

Die Ergebnisse der Literaturrecherche zeigen, dass die dieser Arbeit vorausgehende Annahme „Conclusions-Abschnitte sollten die wichtigsten Schlussfolgerungen enthalten“ nicht weit verbreitet zu sein scheint, obwohl es allgemein anerkannt ist, dass Wissenschaft Wissen erzeugen soll und dass Schlussfolgerungen Wissen sind. Das bedeutet, dass zur Beurteilung der Qualität einer wissenschaftlichen Arbeit die Schlussfolgerungen eines der wichtigsten Indizien sein müssen. Sie sind produziertes Wissen und damit das Ziel der Forschung. Um die Qualität von Artikeln zu beurteilen, ist es nicht ausreichend nur die Conclusions-Abschnitte und dort enthaltene Schlussfolgerungen zu betrachten. Zusätzlich müssten alle anderen Abschnitte betrachtet werden, die Schlussfolgerungen enthalten sollen. Dafür kommen vor allem Diskussion-Abschnitte in Frage, denn eine Form des Diskussion-Abschnitts findet sich in allen betrachteten Artikeln.

In den meisten der gefundenen Artikel wird die Struktur von Artikeln besprochen. Dabei sind Abstract, Einleitung, Methodik, Resultate und schließlich ein Interpretations- Schlussfolgerungs- oder Diskussionsteil wiederkehrend. Selten wird von einem gesonderten Abschnitt am Ende eines Artikels gesprochen, also dem Conclusions-Abschnitt.

2.3 Erkenntnisse aus den Literaturrecherchen

Ziel der Literaturrecherche war es, mich zunächst über das Thema zu informieren, herauszufinden, inwieweit die Annahme gerechtfertigt ist, dass Schlussfolgerungen in Conclusions-Abschnitten enthalten sein sollten und wie dies im SWE gehandhabt wird.

Die anfängliche Literaturrecherche hat mir geholfen, mich mit dem Thema vertraut zu machen, auch wenn es deutlich weniger Artikel gab, als erwartet. Es folgen die Erkenntnisse aus den Literaturrecherchen:

- Meine Annahme über Conclusions-Abschnitte ist gerechtfertigt: Ein Conclusions-Abschnitt ist der letzte Abschnitt eines wissenschaftlichen Artikels und sollte die wichtigsten Schlussfolgerungen zusammenfassen.

- Conclusions-Abschnitte scheinen im SWE wenig Bedeutung zu haben. Von 14 SWE Artikeln sehen nur sechs einen obligatorischen Conclusions-Abschnitt vor. Stattdessen wird immer erwartet, dass die Schlussfolgerungen ausführlich in einem Diskussion-Abschnitt präsentiert werden.

- Im SWE gibt es zu wenig Artikel über das Schreiben und Publizieren von wissenschaftlichen Artikeln. Bücher zu diesem Thema, waren eher allgemeinwissenschaftlicher Natur. Da Conclusions-Abschnitte in jedem Artikel auftauchen, sind sie ein relevantes Thema. Auch Artikel über den gesamten Aufbau und Inhalt wissenschaftlicher Artikel scheinen ein relevantes, aber wenig behandeltes

Thema zu sein. Dies zeigt auch ein Artikel aus dem Jahr 2007, der eine dem Autor sinnvoll erscheinende Lesemethode für wissenschaftliche Artikel beschreibt, die kaum durch Referenzen gestützt wird [43]. Trotzdem wird dieser Artikel, seinen Referenzierungen nach, bis heute als Grundlage für systematische Literaturrecherchen genutzt.

Artikel über das Lesen bzw. Schreiben von wissenschaftlichen Artikeln im SWE sind rar und deuten auf die Tatsache hin, dass sich im SWE zu wenig mit diesem Thema befasst wird. Es bleibt nur die Schlussfolgerung, dass das Thema nie wichtig genug war, um sich damit zu beschäftigen: Möglicherweise, weil SWE traditionell eine Ingenieurwissenschaft ist und in einer solchen hauptsächlich mit Bekanntem Neues konstruiert wird, sei es eine Methodik oder ein Produkt. In diesem Fall ist der letzte Abschnitt immer eine Beschreibung des neu Geschaffenen und seines Nutzens, egal ob der Autor in diesem Abschnitt Schlussfolgerungen präsentieren will oder seine Arbeit zusammenfasst. Da die Schaffung von etwas Neuem Kreativität erfordert, ist die beste Zusammenfassung eine einfache Beschreibung des neu Geschaffenen. SWE ist aber zu einem großen Teil keine reine Ingenieurwissenschaft mehr, sondern hat auch empirische und sozialwissenschaftliche Teile. Möglicherweise fehlt aber aufgrund des Ursprungs des SWE eine Auseinandersetzung mit der Präsentation von Forschung, da diese in den Ingenieurwissenschaften offensichtlich war. Endprodukt Ingeneurwissenschaftlicher Forschung sind Systeme, Techniken oder Veränderungen dieser, daher enthält ein Conclusions-Abschnitt dieser Wissenschaften immer eine Beschreibung des Produktes und des damit verbundenen Nutzens.

- Schlussfolgerungen sind der Wissenschaftstheorie nach Wissen und damit ein essentieller Teil der Forschung. Dies bestätigen auch alle Artikel aus dem SWE, da sie Schlussfolgerungen in Artikeln erwarten.

Wollte man diese Ergebnisse verbessern und erweitern, um ein vollständigeres Bild der Schlussfolgerungen in der Wissenschaft zu erhalten, würde ich folgende Punkte vorschlagen, für die mir in der Ausarbeitung die Zeit fehlte.

- Es könnte nach mehr normativem Gedankengut zu wissenschaftlichen Arbeiten recherchiert werden mit Hilfe eines Wissenschaftstheorethikers. Zusätzlich könnte gezielt nach Wissenschaftstheorie über empirische Arbeiten und Ingenieurswissenschaften gesucht werden. Weiterhin würde es sich anbieten die Delphi-Methode mit entsprechenden Experten, aus z.B. der Wissenschaftstheorie und -rhetorik, durchzuführen. Mit der Delphi-Methode wird das Wissen von Experten in einem systematischen Prozess ergründet [5].

- In der Wissenschaftsrhetorik könnte nach passenden Quellen gesucht werden. Denn Wissenschaftsrhetorik beschäftigt sich mit der Funktion des Überzeugens [81] und ein Artikel soll schließlich von der eigenen Forschung überzeugen. Dies könnte also ein Bereich sein, in dem Struktur und Argumentation von wissenschaftlichen Artikeln diskutiert werden.

- Eine quantifizierende Studie über die Guidelines von SWE- Zeitschriften, um mehr über die aktuellen Erwartungen an wissenschaftliche Artikel oder auch an

konkrete Schlussfolgerungen und Conclusions-Abschnitte zu erfahren. Denn die Strukturen der Artikel werden weniger durch wissenschaftstheoretische Normative als vielmehr durch die Richtlinien der wissenschaftlichen Zeitschriften und die Erwartungen der Leser, die an die einheitliche Form wissenschaftlicher Arbeiten gewöhnt sind, vorgegeben [70].

2.3.1 Argumente für Schlussfolgerungen in Conclusions-Abschnitten

Nach der Wissenschaftstheorie scheint meine Annahme zu Conclusions-Abschnitten und deren Inhalt berechtigt. Dass dies in der Realität nicht umgesetzt wird, war mir aufgrund der Motivation der Studie von Anfang an bewusst, wird aber durch das Fehlen von Artikeln die die Artikelstruktur besprechen, untermauert. Aus den Gesamtergebnissen lässt sich kein eindeutiges Bild konstruieren. Es gibt zu wenig Literatur im Bereich des SWE, die sich konkret mit der Aufbereitung von wissenschaftlichen Artikeln und deren Struktur beschäftigt. Die meisten Artikel, die ich gefunden habe, enthalten keinen Conclusions-Abschnitt, sprechen sich aber auch nicht eindeutig dagegen aus. Zumindest ist die grundsätzliche Wichtigkeit von Schlussfolgerungen in allen Artikeln erkennbar.

Historische Anmerkung: Eine von mir gefundene Quelle, die die gesamte Wissenschaft betrachtet, hat beobachtet, dass nach 1925 die Anzahl der Artikel ohne Conclusions-Abschnitt zugenommen hat. Die Autoren vermuten, dass dies auf die zu dieser Zeit üblich gewordenen Abstracts am Anfang der Artikel zurückzuführen ist, die bereits die wichtigste Schlussfolgerung enthielten. [35]

Führt man diesen Gedanken weiter, so kann man zu dem Schluss kommen, dass zumindest um 1925 die Erwartung der wissenschaftlichen Gemeinschaft an Conclusions-Abschnitte die gewesen sein muss, dass ein Conclusions-Abschnitt die wichtigsten Schlussfolgerungen des Artikels enthält. Da die wichtigste Schlussfolgerung bereits im Abstract enthalten ist, scheint der Conclusions-Abschnitt seine Funktion verloren zu haben.

Mögliche Begründungen für Schlussfolgerungen in Conclusions-Abschnitten und die Existenz solcher Abschnitte fasse ich im Folgenden zusammen. Nicht alle Begründungen sind durch Quellen belegt, einige beruhen auf Annahmen, die mir einleuchtend erscheinen.

1. **Wissensgenerierung effizient halten:** Neues Wissen baut auf bekanntem Wissen auf, und Wissen besteht aus Schlussfolgerungen und nicht aus Ergebnissen. Da es die Aufgabe der Wissenschaft ist, neues Wissen zu generieren, sollte die Forschung das neue Wissen so effizient wie möglich verbreiten. Es sollte in verständlicher Form als konkrete Schlussfolgerung festgehalten werden und nicht nur in Form von Ergebnissen, die eine Schlussfolgerung erlauben. Denn in diesem Fall obliegt es jedem einzelnen Empfänger, die letzte kognitive Arbeit zu leisten, um die Ergebnisse in Wissen umzuwandeln. Das so generierte Wissen basiert immer auf unvollständigen Daten. Außerdem würde der Aufwand, die Ergebnisse in Wissen umzuwandeln, nicht nur einmalig vom Autor eines Artikels betrieben werden müssen, sondern von jedem einzelnen Leser. Wenn ein

Forscher mehrere Artikel liest, wird er wahrscheinlich damit beschäftigt sein, die Ergebnisse der Artikel, die er gelesen hat, in Schlussfolgerungen umzuformulieren, anstatt eigene Hypothesen auf der Grundlage des erworbenen Wissens aufzustellen. Schlussfolgerungen sind daher wichtig und besonders wertvoll, wenn sie ausformuliert und nicht nur durch Ergebnisse impliziert werden.

Der Wissenserwerb durch die Lektüre von Artikeln könnte durch die Auswahl und Wiederholung der wichtigsten Erkenntnisse erleichtert werden. Das könnte noch einprägsamer und effizienter gestaltet werden, in dem in jedem wissenschaftlichen Artikel ein Abschnitt vorgesehen wird, der genau diesem Zweck gewidmet wäre. Dies würde der Idee entsprechen, dass wissenschaftliche Artikel einer festen Struktur folgen und damit die Nutzung solcher Artikel effizienter machen [35, 59, 46].

2. **Der letzte Abschnitt eines Artikels:** Allein aufgrund der Positionierung des Conclusions-Abschnitts und der Natur der Wissenschaft, sollten im Conclusions-Abschnitt die wichtigsten Schlussfolgerungen eines Artikels stehen. Der Zweck wissenschaftlicher Artikel ist es, Wissen glaubwürdig zu teilen und zu präsentieren. Damit ähnelt der Artikel einer argumentativen Rede in der üblicherweise am Ende auf die Ausgangsfrage Bezug genommen wird. Im Fall von Forschung wäre das eine Beantwortung der Forschungsfrage, was die Schlussfolgerungen aus der Forschung sein sollten. Weiterhin scheint der Schlussteil jeder Rede und jedes Artikels eine geeignete Gelegenheit zu sein, zu wiederholen, was dem Leser am dringendsten vermittelt werden soll. Zum einen wird so dem Leser etwas doppelt vorgetragen und bleibt so besser im Gedächtnis, zum anderen wird das Wichtigste vom Rest getrennt, hebt sich so besser ab und erscheint relevanter.

3. **Persönliche Leseerwartungen:** Wissenschaftliche Artikel haben eine recht einheitliche Struktur [35, 46, 59]. Diese einheitliche Struktur scheint sich vor allem aufgrund der Erwartungen von Forschern aufrechtzuerhalten [35]. Die Struktur erleichtert den Forschern die Navigation innerhalb der Artikel und ermöglicht ein effizienteres Lesen. Obwohl die in der Literatur übliche Gliederung der Artikel (Einleitung, Methodik, Resultate, Diskussion) keinen Conclusions-Abschnitt vorsieht [35, 46], beziehen sich Leseempfehlungen für wissenschaftliche Artikel auf einen Conclusions-Abschnitt [84, 43]: Um festzustellen, ob ein Artikel in den persönlichen Lesehorizont passt, wird empfohlen, zunächst den Titel, dann das Abstract und schließlich die Conclusions zu lesen. Der Titel lasse das Themenfeld bestimmen, das Abstract fasse grob zusammen und gebe Forschungsziele an, der Conclusions-Abschnitt beantworte die Forschungsfragen. Mit den Antworten auf die Forschungsfragen sollte jeder Leser das genaue Thema der Arbeit begreifen können. Fehlt der Conclusions-Abschnitt muss zur initialen Evaluierung die Diskussion gelesen werden, was aufgrund der Ausführlichkeit der Diskussion mehr Zeit in Anspruch nimmt und somit keinen schnellen Einblick verschafft.

4. **Implikationen aus den Artikeln:** Eine Überschrift sollte immer auf den nachfolgenden Inhalt Bezug nehmen, idealerweise fasst sie diesen zusammen. Die meisten Conclusions-Abschnitte tragen die Überschrift „Conclusion“ oder „Conclu-

sions“ – ersteres bedeutet Schlussfolgerung im Singular oder auch Abschluss, zweiteres bedeutet übersetzt Schlussfolgerungen. Die Überschrift „Conclusions“ kündigt daher folglich Schlussfolgerungen an. Auch wenn nur ein Abschluss angekündigt wird, kann im Kontext wissenschaftlicher Arbeiten eine Schlussfolgerung erwartet werden, wie im zuvor genannten Argument zum letzten Abschnitt einer Schrift diskutiert.

Einige Autoren wecken die Erwartung von Schlussfolgerungen in ihrem Conclusions-Abschnitt, indem sie dies in ihrer Gliederung am Ende der Einleitung versprechen aber nicht einlösen, indem sie den Abschnitt ihres Artikels mit Varianten der Formulierung „we end our article with conclusions“ beschreiben, aber trotzdem keine Schlussfolgerungen im letzten Abschnitte nennen. So eine uneingelöste, aber ausformulierte Versprechung zeugt davon, dass manche Autoren anscheinend nicht über ihren eignen Text nachdenken oder „Schlussfolgerungen“ grundsätzlich anders verstehen, als es in der Wissenschaft üblich sein sollte.

3 Entwicklung eines Codebuchs

In diesem Kapitel werde ich genutzte Methoden zur Erstellung eines Codebuchs und anschließend meine Forschungsarbeit dazu in grober chronologischer Abfolge beschreiben. Der Fokus liegt auf meinem Arbeitsprozess.

3.1 Methoden

Für die **Datenanalyse und Theorieentwicklung** habe ich Techniken und Konzepte der „Grounded Theory Methodology“ (GTM) angewandt. Diese Techniken werden zuerst erläutert. Es folgt eine Zusammenfassung weiterer persönlicher Organisationsmethoden und genutzten Tools, wie z.B. MAXQDA.

3.1.1 Angewandte Techniken

„Grounded Theory Methodology“ heißt auf deutsch gegenstandsbezogene Theoriebildung und ist eine qualitative Methode zur Datenanalyse. Ziel der Methode ist es, aus den Daten heraus eine Theorie bzw. miteinander zusammenhängende Theorien zu entwickeln [13, 36, 32]. Das bedeutet, dass bei dieser Methode die Daten am Anfang stehen und die Theorie induktiv aus den Daten abgeleitet wird [13]. Das wird durch ein sehr gutes Verständnis der Daten ermöglicht; zu Beginn der Forschung entwickeln Forschende Hypothesen aus den vorhandenen Daten und verfeinern diese immer weiter, sobald neue Daten hinzukommen, so dass die entstandenen Theorien am Ende der Forschung weiterhin alle vorhandenen Daten genau beschreiben [13].

Die Verwendung von Techniken aus der GTM als Analysemethode für diese Arbeit ist passend, da die Motivation für die Studie aus den Daten kam und es möglich werden soll Faktoren zu finden, die die Qualität genau dieser Daten beurteilen. Ausgangspunkt war ein unbefriedigendes Gefühl, das sich oft nach der Lektüre von SWE-Artikeln einstellte. Ursachen für dieses Gefühl zu ergründen ist Forschungsthema dieser Arbeit. Ein Ansatz von den Daten zur Hypothese hin, wie in der GTM

genutzt, eignet sich um den Ursprung des unbefriedigten Gefühls zu erforschen. GTM ist außerdem eine nützliche Methode, um die soziologischen und menschlichen Aspekte von SWE zu untersuchen [36]. In diese Richtung geht die Ergründung der Conclusions-Abschnitte, da wissenschaftlichen Artikel Kommunikationsmittel sind.

Bevor ich die einzelnen genutzten Techniken der GTM beschreibe, möchte ich einen **Überblick über die GTM** geben: Es gibt drei verschiedene Schulen der GTM, die unterschiedliche Schwerpunkte setzen: Nach *Glaser*[32] gibt es genau eine Realität, zu der jeder Forscher kommen wird, wenn er sich intensiv genug mit den Daten auseinandersetzt. *Strauss und Corbin* [15] gehen davon aus, dass es eine Realität gibt, erwarten aber nicht, dass jeder diese auf die gleiche Weise wahrnimmt. Damit Forscher bei der Anwendung der GTM so wenig wie möglich ihren eigenen Vorurteilen und Verzerrungen unterliegen, haben Strauss und Corbin systematische Verfahren eingeführt. *Charmaz* ist der Meinung, dass es nicht die eine Realität gibt, sondern viele verschiedene Realitäten, je nach Perspektive. Ihre GTM legt daher großen Wert darauf, dass die Forscher während ihrer Arbeit möglichst viele verschiedene Perspektiven einnehmen. [74]

Alle drei Varianten der GTM haben grundsätzlich einen ähnlichen Ablauf.

Methodisch beginnt die Arbeit mit der GTM ohne ausführliche Literaturrecherche [79, 19] und ohne die Formulierung einer konkreten Hypothese. Stattdessen wird ein Interessenfokus definiert, um zu verhindern, dass sich Forscher zu sehr auf bereits bestehende Hypothesen konzentrieren [36, 79, 19]. Strauss und Corbin empfehlen, abweichend von den anderen GTM-Schulen, eine schmale anfängliche Literaturrecherche, weisen aber auch darauf hin, dass diese sowohl hilfreich als auch hinderlich sein kann[15]. Durch meine anfängliche Literaturrecherche konnte ich ein grundlegendes Verständnis aufbauen.

Nach der optionalen Literaturrecherche folgt die Datenerhebung, in dieser Arbeit lagen die Daten bereits vor (3.2).

Sobald Daten vorliegen, beginnt der Analyseteil mit der Entwicklung eines Codesystems durch *offene Codierung*. Dabei markiert die forschende Person verschiedene Informationen unterschiedlich (z.B. Inhalt, Audrucksweise, Emotionen, ...) und kategorisieren damit Informationen im Test. Diese Kategorien werden Codes genannt. Während des offenen Codierens werden alle Auffälligkeiten festgehalten und Codes und codierte Textstellen bzw. Segmente werden miteinander verglichen, um die Codes besser zu verstehen. Codes können Ober- und Unterkategorien haben und verschiedene Dimensionen, die den codierten Inhalt genauer spezifizieren können [32]. Auf diese Weise wird ein Codesystem entwickelt, das es ermöglicht, die Daten zu ordnen und Zusammenhänge leichter zu erkennen. Die Codes werden in einem Codebuch festgehalten, welches das Codesystem beschreibt.

Im weiteren Verlauf der Studie werden weitere Daten mit Hilfe des Codebuchs codiert und analysiert, bis das Codebuch abgeschlossen ist. Je mehr Daten analysiert werden, desto passendere Theorien können entstehen [32]. Das offene Codieren ist eine Analysetechnik und damit Teil der Datenanalyse, da dabei Konzepte erkannt werden können. Während des Forschungsprozesses können nicht nur Codes, sondern auch Konzeptideen, vage Gedanken und Ahnungen aufkommen und sollten möglichst sofort in Memos festgehalten werden und im Verlauf der Forschung ständig weiter genutzt und verglichen werden. Diese Technik heißt *Memoing*. Memos können

Definitionen von Codes sein, Beziehungen zwischen Codes beschreiben oder Interpretationen der Daten sein. Memos dürfen auch unfertige Ideen und Gedanken sein [36, 32].

Während der Datenanalyse werden, wie Memos, auch Codes, codierte Textstellen und Theorien ständig weiter miteinander verglichen und betrachtet. Durch das ständige Vergleichen sollen Zusammenhänge sichtbar werden und Hypothesen entstehen bzw. klarer werden. Genauso wie Hypothesen können weitere Codes entstehen oder verfeinert werden. Es wird so lange verglichen, bis die entstandenen Hypothesen und Konzepte gesättigt sind. Dies zu bemerken liegt in der Verantwortung und im Gefühl des Forschers und wird als *theoretische Sättigung* bezeichnet. Die Fähigkeit des Forschers, Konzepte und Theorien in den Daten zu erkennen, wird als *theoretische Sensitivität* bezeichnet und soll bei Strauss und Corbin durch systematische Analysetechniken vereinfacht werden. [15, 36, 79]

Darüber hinaus werden während der gesamten Datenanalyse weiterhin Daten erhoben und zwar immer dann, wenn der Forscher feststellt, dass es Lücken in den Daten gibt. Für meine Forschung war es nicht notwendig, weitere Daten zu erheben. Allerdings habe ich mitunter Artikel nach bestimmten Merkmalen ausgewählt, z.B. nach schlechten Schlussfolgerungen, wenn ich mehr Beispiele dafür benötigte [36, 79].

Während der Arbeit an meiner Thesis habe ich auf Anraten meines Betreuers die GTM-Variante von Strauss und Corbin genutzt, da die dort enthaltenen systematischen Verfahren das Erlernen der Methode erleichtern. Das Buch von Strauss und Corbin ist mit Beispielen ausgestattet, die die Technik des offenen Codierens und GTM im Allgemeinen leicht verständlich erscheinen lassen. Die GTM-Variante von Strauss und Corbin enthält viele systematische Ansätze, die für mich ein willkommener, gut definierter Startpunkt waren.

Für meine Forschung habe ich das Konzept der *theoretischen Sensitivität* genutzt. Theoretische Sensitivität ist die Fähigkeit entscheidende Stellen in der großen Menge an Daten zu erkennen, um passender Theorien zu entwickeln [15]. Neben der Fähigkeit, Relevantes erkennen zu können, ist es wichtig, diese Fähigkeit während der Forschung gezielt zu verbessern. Dafür habe ich, nach Strauss und Corbin folgende Techniken angewandt: Offenes Codieren, Weitercodieren von mehr Daten, Memoing in Form von Memos und einem Forschungstagebuch, Microcoding und Vergleichen von Extremen.

3.1.2 Sonstiges

Um offenes Codieren und Memoing zu vereinfachen, habe ich MAXQDA verwendet, ein Programm, das speziell für qualitative Forschung mit Fokus aufs Codieren entwickelt wurde. In diesem Tool können Texte codiert werden. Codesysteme und Memos können erstellt, sortiert und miteinander und zu Textsegmenten verlinkt werden. Außerdem ist es möglich, Codes oder codierte Segmente auf verschiedene Arten zu vergleichen oder textuell nach gemeinsamen Wörtern zu analysieren.

Da die Memos innerhalb von MAXQDA nicht immer ausreichten, um meine Gedanken festzuhalten, habe ich während meiner Recherchephase zusätzlich Miro (`https://miro.com/de/`), Overleaf(`https://de.overleaf.com`) und handschriftliche Notizen

auf Papier verwendet. Letzteres ist selbsterklärend und wurde von mir hauptsächlich zum Sammeln von Tages- oder Wochenzielen verwendet. Miro ist eine webbasierte Whiteboard-Anwendung, die ich während meines letzten Studienabschnitts häufig genutzt habe, um erworbenes Wissen geordnet festzuhalten. Im Rahmen meiner Arbeit verwendete ich sie, um Codes zu organisieren und Erkenntnisse aus meiner Forschung systematisch nebeneinander zu stellen. Zu Beginn meiner Arbeit habe ich Miro auch als Forschungstagebuch verwendet, aber mit der Zeit wurden meine Tagebucheinträge immer länger, so dass ich auf Overleaf wechselte und dort ein ausführlicheres Forschungstagebuch geführt habe. Overleaf ist ein web- und LaTeX-basierter Texteditor. Diesen habe ich auch für das Schreiben dieser Arbeit verwendet. Gegen Ende der Arbeit, nachdem ich einige nachteilige Gewohnheiten bei mir erkannt hatte, habe ich auch angefangen, bestimmte Dinge auf Post-Its zu schreiben, damit sie präsenter für mich waren. Mehr dazu unter Selbstreflexion (siehe 6.1).

3.2 Datenbasis

Für meine Forschung musste ich keine Daten erheben. Da ich die Datensammlung der qabstracts-Studie nutzen konnte.

Die Datengrundlage umfasst Artikel der fünf wichtigsten SWE-Zeitschriften aus dem Jahr 2022: „Empirical Software Engineering“, „International Conference on Software Engineering“, „Information and Software Technology“, „Transactions on Software Engineering and Methodology“ (TOSEM) und „Transactions on Software Engineering“. Von jeder Zeitschrift wurde die gleiche Anzahl Publikationen zufällig ausgewählt, mit Ausnahme der TOSEM, welche dafür nicht genug Artikel veröffentlicht hatte.

Für diese Arbeit wurden nur Artikel verwendet, bei denen es sich nicht um Artikel handelt, in denen Neues entwickelt wird. Solche Artikel, die Neues entwickeln, sind Design-Artikel. Die verwendeten Artikel sollten empirischer Natur sein und keine neuen Methoden oder neue Implementierungen vorstellen. Ob es sich um einen Design-Artikel handelt oder nicht, lässt sich bereits im Abstract erkennen. In der qabstracts-Studie wurden solche Design-Sätze mit Codes kenntlich gemacht, so dass ich nach diesem Code Artikel auswählen konnte. Daher konnte ich mir die bereits codierten Artikel der qabstracts-Studie zunutze machen, um Design-Artikel auszusortieren.

Die in dieser Studie analysierten Artikel wurden zu drei Viertel willkürlich und nach Verfügbarkeit ausgewählt, das letzte Vietel wurde ausgewählt, indem innerhalb von Conclusions-Abschnitten nach subjektiv besonders gut oder besonders schlecht erscheinenden Sätzen gesucht wurde.

Die vollständige Liste der von mir codierten Artikel, lässt sich unter folgendem Link finden:

`https://git.imp.fu-berlin.de/ailis/ma-thesis-quality-of-conclusions`.

3.3 Forschungsarbeitsbericht zur Entwicklung eines Codebuchs

In diesem Teil meiner Arbeit lag der Hauptaufwand darin, qualitative Datenanalyse zu betreiben. Das Ziel, auf das ich hinarbeitete, war die Erstellung eines Codebuchs,

das eine quantitative Studie zur Bewertung der Qualität der Schlussfolgerungen ermöglichen würde. Eigentliches Ziel der Forschung war es, mehr darüber herauszufinden, warum Conclusions-Abschnitte unbefriedigend sein können. Aus diesen Gründen habe ich mich in dieser Phase hauptsächlich mit offenem Codieren beschäftigt.

Diese Zeit mit Fokus auf die Entwicklung eines Codebuchs, lässt sich grob in zwei Phasen einteilen. In der ersten Phase ging es darum ein Verständnis für Conclusions-Abschnitten zu erlangen. In der zweiten Phase ging es darum verschiedene Arten von Conclusions-Abschnitten zu finden. Für jede der Phasen wird folgend die Umsetzung beschrieben, dann Ergebnisse zusammengefasst und abschließend der Forschungsprozess reflektiert.

3.3.1 Lernphase

Ende 2022 begann ich, mich über qualitative Forschung mit Fokus auf GTM zu informieren. Techniken zu erlernen und zu verstehen, fiel mir allerdings ohne ein vollständiges Beispiel oder konkrete Möglichkeiten in der Praxis schwer. Die erste Phase des offenen Codierens diente mir somit als Lernphase. Nach einer ersten kurzen Einführung in Schlussfolgerungen und Conclusions-Abschnitte in Form einer Literaturrecherche, habe ich mich erneut mit Techniken qualitativer Datenanylse vertraut gemacht. Dazu habe ich SWE-Forschungsartikel gelesen, die GTM als Methode verwendet haben. Mein Verständnis von passenden Techniken qualitativer Datenanalyse, wie z.B. offenes Codieren, Microcoding und Memoing, habe ich in Miro festgehalten.

Meine erste Codierungs-Sitzung begann mit Papier, Stift, Marker und ausgedrucktem Text, noch bevor ich das Buch von Strauss und Corbin gelesen hatte. Diese Herangehensweise bot nicht genug Platz, um ordentlich zu arbeiten.

Meine zweite Codierungs-Sitzung fand mit Hilfe von MAXQDA statt und mit mehr Wissen über GTM-Techniken von Strauss und Corbin. Dieser Versuch verlief daher deutlich strukturierter, war aber noch immer sehr anspruchsvoll. Microcoding anzuwenden, fiel mir schwer. Ich benötigte viele Stunden, um die ersten Artikel zu codieren. Dabei entstanden nur sehr grobe Codes, die entweder sehr Offensichtliches beschrieben oder sehr Spekulatives. Mit dem Thema des Artikels hatte ich mich bisher noch nicht befasst, sodass es herausfordernd war den Abschnitt zu verstehen. Daher habe ich mich an den Satzintensionen orientiert und häufig vorkommende Wörter näher betrachtet. Für den ersten vollständig codierten Conclusions-Abschnitt benötigte ich daher zwei ganze Tage. Danach hatte ich die grundlegenden Codierungs-Funktionen des Tools MAXQDA verstanden, aber noch kein gutes Verständnis von Conclusions-Abschnitten oder dem Codieren erlangt. Für den nächsten Conclusions-Abschnitt benötigte ich nur noch einen Tag und den dritten Abschnitt konnte ich innerhalb von drei Stunden bewältigen. Die Zunahme an Geschwindigkeit beim Codieren erkläre ich mir durch die möglich gewordene Nutzung von entstandenen Codes und dem erlernten Umgang mit MAXQDA.

In dieser initialen Codierungsphase hatten die Codes oft sehr lange Namen und es kam noch häufig zu Doppelungen von Codes für dieselbe Information. Es war mir noch nicht möglich, für meine Forschung relevante Textstellen zu erkennen, da ich noch damit beschäftigt war, mich überhaupt in den Texten zurecht zu finden. Das Anwenden der Technik des Microcoding fiel mir schwer, vor allem das Beleuchten

einzelner Wörter brauchte viel Zeit. Deshalb habe ich vorübergehend davon Abstand genommen, Microcoding zu verwenden und bin zu intuitiven Codierungen ohne Granularität übergegangen – mit der Hoffnung, ein besseres allgemeines Verständnis für Schlussfolgerungen und Conclusions-Abschnitte zu entwickeln.

Nach zwei Wochen des Codierens begann ich erstmals damit, Codes umfassend miteinander zu vergleichen und nicht nur selektiv nach Bedarf (siehe beispielhaft Abbildung 5 auf Seite 91). Beim Vergleichen habe ich Codes zusammengelegt, gelöscht oder als verwandt eingestuft. Am einfachsten war es für mich, darüber nachzudenken, ob Codes zusammengelegt werden sollten oder nicht. Legte ich Codes nicht zusammen, obwohl sie sehr ähnlich waren, fiel es mir schwer zu beschreiben, was der entscheidende Unterschied war. Zu diesem Zeitpunkt hatte ich noch nicht realisiert, dass der nächste Schritt darin bestand, fast gleiche Codes näher zu untersuchen und zu konkretisieren warum sie nicht gleich waren. Das führte zu einer Menge an fast gleichen Codes, die nicht ausdefiniert waren. Es erschien mir so, als ob ich noch kein aureichends Verständnis von Schlussfolgerungen hatte, um diese zu Analysieren und Auszudefinieren. Daher fokussierte ich mich darauf, weitere Conclusions-Abschnitte zu codieren und betrachtete die dort enthaltenen Schlussfolgerungen näher.

Währenddessen erkannte ich, dass ich noch nicht in der Lage war, Schlussfolgerungssätze zu erkennen und von andersartigen Sätzen zu unterscheiden. Zum Beispiel waren Resultate und Schlussfolgerungen mitunter schwer zu unterscheiden, z.B. wenn ähnliche ähnliche Satzanfänge verwendet wurden. Eine weitere Erkenntnis war, dass es Abschnitte mit Schlussfolgerungen gibt, in denen die Schlussfolgerungen ausformuliert und dem Leser erklärt werden und andere Abschnitte mit Schlussfolgerungen, in denen Informationen und Beobachtungen angeführt werden, aber das Schlussfolgern den Lesenden selber überlassen wird. Letzteres ist nach meiner Definition von Conclusion-Abschnitten nicht erwünscht. Ein Beispiel für beide beschriebenen Fälle ist folgender Satz, auf den kein weiterer Schlussfolgerungssatz im Originaltext folgt: „We analyzed a total of 276 projects, and found that 85 of them present one or more rotten green tests, accounting for a total of 665 rotten green tests“ [2]. Dies ist ein Resultat, keine Schlussfolgerung, da dieser Satz nur Fakten nennt, die durch die Studie entstanden sind. Die Phrase „*we found*“ kann auch als Einleitung für Schlussfolgerungssätze genutzt werden.

Nach weiteren Codierungen von Artikeln und weiteren Vergleichen von Codes und codierten Segmenten betrachtete ich erstmals genau die Codes der qabstracts-Studie. Dies hatte ich zuvor aus demselben Grund vermieden, weswegen initiale Literaturrecherchen in der GTM vermieden werden: Ich wollte verhindern, dass ich von den Codes der qabstracts-Studie zu sehr beeinflusst werden würde. Die meisten der Codes der qabstracts-Studie sind Codes, die den strukturellen Inhalt des Satzes innerhalb des Abstracts beschreiben. Meine Codes die den strukturellen Inhalt beschrieben ließen sich in ausdefinierter Form in der qabstracts-Studie wieder finden, so dass ich einige dieser Codes übernahm. Insbesondere die Bezeichnungen der Codes aus der qabstracts-Studie waren prägnant und passend, anders als meine Code-Namen. Ein Beispiel dafür ist der Code objective, den ich bei mir goal, purpose of study oder what to find out genannt hatte. Andere Code-Konzepte waren mir neu, beispielsweise die Aufteilung der zukunftsorientierten Sätze. Diese sind in der qabstracts-Studie

nach der Ausdrucksweise differenziert.

Nach zwei Monaten des Codierens fiel es mir zum ersten Mal leicht, einen Artikel zu codieren. Zu diesem Zeitpunkt hatte ich bereits einige Codes erstellt, verglichen, fusioniert und auch gelöscht. Codes sind in meiner Arbeit gleichzustellen mit Kategorien die verschiedenartige Textteile beschreiben. Ich hatte ein grobes Verständnis von Conclusions-Abschnitten erlangt und konnte erkennen, welche Sätze welchen strukturellen Inhalt zum Abschnitt beitragen. Das bedeutet, dass ich mich auch ohne Fachwissen in Conclusions-Abschnitten zurechtfand. Damit war ich in der Lage, mich mehr auf Details in Formulierungen zu fokussieren, wie z.B. Zeitformen.

Bisher hatte ich mich unbewusst stark auf die Struktur konzentriert, die oft leichter zu erkennen war als Formulierungsfeinheiten und zu objektiven Codes führte. Daneben entstanden aber auch Codes, die subjektive Vermutungen über mögliche Motive der Autoren waren, die sich mehr mit der Frage beschäftigten, warum die Autoren bestimmte Sätze geschrieben hatten. Das Folgern von vermuteten Gründen von Sätzen auf generelle Einstellungen der Autoren zum gesamten Conclusions-Abschnitt oder sogar auf ihr Wesen erschien mir schnell als zu spekulativ. Denn wissenschaftliche Artikel sind keine spontanen unveränderten Ausdrucksweisen, sondern stark genormte und in einem iterativ Prozess verbesserte Formulierungen. Die Normen stammen sowohl aus konkreten Vorgaben von einzelnen Zeitschriften, als auch aus unterbewusst gebildeten Standards durch das Lesen anderer wissenschaftlicher Artikel. Daher wandte ich mich möglichst objektiven Beurteilungen von Sätzen zu und versuchte, aus diesen möglichst viele Informationen abzuleiten, die mir Aufschluss über die Qualität von Conclusions-Abschnitten geben könnten.

Nach drei Monaten offenem Codieren hatte ich Schlussfolgerungen begriffen, aber noch nicht verinnerlicht. Obwohl ich eine einleuchtende Definition hatte, fiel es mir weiterhin schwer, Schlussfolgerungen sicher zu codieren. Ich versuchte dies zu verbessern, indem ich potentielle Schlussfolgerungssätze mit meinem Betreuer diskutierte und nach weiteren, bereits codierten Beispielen in der qabstracts-Studie suchte. Letzteres eröffnete mir, dass meine Sichtweise auf Schlussfolgerungen noch nicht ausgereift war und eher zu eng schien. Einige der Beispiele schienen mir unter keinen Umständen Schlussfolgerungen zu sein, und bei anderen verstand ich die Differenzierungen nicht: Besonders schwer war es, zwischen Existenzbeweisen und ausgelassenen Verallgemeinerungen zu unterscheiden. Beispielsweise ist dies bei folgendem Satz unklar: „Our detectors indicate that CIPs can be identified using static analysis“ [30]. Um herauszufinden, ob diese Aussage ein Existenzbeweis ist oder eine schlecht formulierte Verallgemeinerung, müsste in den Hauptteil der Studie geschaut werden. Das veranschaulicht, dass es nicht immer einfach ist Schlussfolgerungen ohne Wissen über die Forschung zu beurteilen. Dies zeigt wiederum, dass das Lesen von Conclusions-Abschnitten ohne das Lesen des Hauptteils der Studie nicht immer möglich ist, auch wenn dem so sein sollte [43, 57].

Dadurch wurde ich sicher beim Codieren aller Codes, die konzeptionell weit von Schlussfolgerungen entfernt waren. Ich benötigte mehr Zeit zum codieren möglicher Schlussfolgerungssätze. Da ich noch keine Theorien gefunden hatte, begann ich unbewusst, mich auf die nicht schlussfolgernden Codes zu konzentrieren und in diesen neue Dimensionen und verschiedene Ausdrucksweisen zu erforschen, um zumindest einen Teil des Codebuchs fertigstellen zu können. Ich erhoffte mir außerdem, beim

Erforschen der restlichen Codes mit so vielen weiteren Schlussfolgerungen in Berührung zu kommen, dass sich meine theoretische Sensitivität erhöhen würde und sich ein besseres Verständnis der Schlussfolgerungen einstellen würde.

Zusammenfassung – Diskussion Am Ende dieser Phase, Ende April, hatte ich eine Literaturrecherche durchgeführt, meine ersten Artikel codiert, das Codieren in MAXQDA erlernt, mit dem Codebuch begonnen, ein grobes Verständnis vom Inhalt der Schlussfolgerungen erlangt und eine ungefähre Vorstellung davon, was Schlussfolgerungen sein könnten.

Die Codes, die in dieser ersten Phase entwickelt wurden, sind ein Produkt des Lernens und des Ausprobierens, sie waren noch nicht ausreichend voneinander abgegrenzt und nicht alle Definitionen waren eindeutig. Die initial entwickelten Codes haben viel zu meinem Lernprozess und meiner Fähigkeit zu codieren beigetragen, aber nur wenige Codes aus dieser Phase sind in der letzten Version des Codebuchs enthalten.

Nachdem ich anfangs die systematischen Analysemethoden von Strauss und Corbin eher mühsam und erfolglos angewendet hatte, habe ich diese nicht mehr verwendet. Vielmehr habe ich mich darauf konzentriert, den Inhalt des Textes mit Hilfe meiner ersten Codes zu verstehen, um vor allem die Schlussfolgerungen zu differenzieren. Dies hat letztlich dazu geführt, dass ich erst am Ende meiner Arbeit zu der Erkenntnis gekommen bin, dass es sinnvoll ist, den Text bewusst aus verschiedenen Perspektiven zu betrachten, da nicht alle Perspektiven gleichermaßen relevant sind.

Ich habe außerdem gelernt, dass man die Schlussfolgerungen nicht immer mit einem guten Gefühl kategorisieren kann, wenn man nur die Conclusions-Abschnitte betrachtet. Beispielsweise ist manchmal unklar, ob ein Existenzbeweis wirklich ein Beweis ist oder ob es sich nur um ein Ergebnis und damit um eine unterlassene Verallgemeinerung handelt. Unter anderem führte auch das dazu, dass ich mich regelrecht in das Verständnis von Schlussfolgerungen verbissen habe. Denn ohne die Fähigkeit, meinen Forschungsgegenstand zu erkennen, schien es mir unmöglich, etwas über ihn herauszufinden.

Selbstreflexion: Wenn ich noch einmal qualitative Forschung mit dem Konzept der theoretischen Sensitivität lernen würde, würde ich weniger betonen, dass der Prozess eine lange Durststrecke sein kann, und diesen stattdessen eher wie einen langen unbekannten Weg beschreiben. Außerdem würde ich darauf achten, mehr Microcoding zu verwenden, regelmäßige Pausen zum Zusammenfassen meiner bisherigen Arbeit einzulegen und mir Zeit zu nehmen, um Ziele zu setzen, zu prüfen und ggf. zu ändern.

Eine lange **Durststrecke** signalisiert für mich eher ein unbeteiligt sein der eigenen Person an einer Situation und der Veränderung der Situation. Damit einhergehend scheint mir bei einer Durststrecke eine Planung des Vorgehens zur Verbesserung der Lage unwichtig. Ich bin mir unsicher, warum mir diese Formulierung während meiner Arbeit so sehr präsent war. Aber das hat mich persönlich dazu verleitet, weitere Quellen mit meinen bereits bekannten Codes anzuschauen, anstatt auffällige Textstellen innerhalb der alten Dokumente mit den Methoden von Strauss und Corbin genauer zu betrachten. Ich habe es auch vernachlässigt, in regelmäßigen Abständen

einen Schritt zurückzutreten und über mein Vorgehen nachzudenken. Ich vertraute darauf, dass ich irgendwann beim Lesen, Codieren und Vergleichen dieser Quellen etwas finden würde, dass mir zuvor nicht aufgefallen war. Rückblickend lässt sich sagen, dass die Realisierung und bewusste Auseinandersetzung mit verschiedenen Betrachtungsebenen (z.B. struktureller Inhalt, Rhetorik, Bestimmtheit von Aussagen, Abstraktionsebene, ...) hilfreich gewesen wäre. Im Laufe meiner Arbeit habe ich mich mit verschiedenen Ebenen auseinandergesetzt, aber immer nur punktuell und intuitiv, wenn etwas offensichtlich wurde. Und selbst dann habe ich nicht daran gedacht, die gefundene Perspektive oder Dimension auf andere Codes anzuwenden.

Microcoding war der Oberbegriff für einige systematische Codierungsansätze von Strauss und Corbin. Es war schwierig für mich, diese Techniken anzuwenden. Zunächst, weil ich mit nichts vertraut war, weder mit dem Prozess des Codierens, noch mit dem Inhalt, den ich codierte, noch mit der Software, die ich dafür benutzte. Was mir die Anwendung der Technik zusätzlich erschwerte, ohne dass ich es gleich bemerkte, war, dass die Beispiele, die die Techniken bildlich erklärten, immer solche waren, die versuchten, den Inhalt des zu codierenden Textes besser zu verstehen und darin Facetten zu entdecken. Was ich aber untersuchen wollte, war nicht der Inhalt und dessen Bedeutung, sondern die Ausdrucksweise der Autoren, also die Rhetorik und deren Bedeutung. Erschwerend kam hinzu, dass ich mit den Conclusions-Abschnitten aus wissenschaftlichen Publikationen keine natürliche Ausdrucksform von Menschen in Textform vor mir hatte, sondern eine stark komprimierte Form. Diese Form ist angepasst an die Erwartungen der Wissenschaft als Geschäft und Prozess. Dies steht z.B. im Gegensatz zu ausführlichen Interviews, in denen Menschen über ihre Erfahrungen berichten. Hier kann man aus der Ausdrucksweise Rückschlüsse auf die Person selbst ziehen. Theoretisch ist dies auch bei Conclusions-Abschnitten möglich, aber die Gelegenheiten dazu sind wesentlich geringer. Spontane, unreflektierte Äußerungen, die in Interviews möglich sind, oder auch Denkpausen fallen in mehrfach Korrektur gelesenen wissenschaftlichen Artikeln weg. Rückblickend hätte ich die systematischen Codierungsansätze von Corbin und Strauss systematischer nutzen sollen, indem ich diese konkret auf interessante Textstellen angewandt hätte. Dies hätte mir die Möglichkeit gegeben, die Techniken mehr zu nutzen und damit systematischer und nicht ausschließlich intuitiv zu arbeiten. So hätte ich auch die Möglichkeit gehabt, mir Tagesziele zu setzen, die erreichbar sind, z.B. das Microcoding für jede Perspektive einmal durchzuarbeiten.

Ich habe festgestellt, dass das **Setzen von Zielen** in erreichbaren Schritten für mich persönlich sehr wichtig ist, aber das Finden von Zwischenzielen beim Erstellen eines Codebuchs, dem offenen Codieren und ständigem Vergleichen von Daten, Codes und Memos nicht immer einfach ist. Ohne Zwischenziele gelang es mir nur schwer weiterzuarbeiten. An den wenigen Tagen, an denen ich mir Aufgaben stellen konnte, weil Sprechstunden anstanden oder ich etwas nachlesen wollte, war ich konzentrierter, motivierter und dadurch deutlich länger produktiv.

Ziele zu setzen fiel mir aus mehreren Gründen schwer. Ein Grund war meine zeitweilig unpassende Vorstellung des Ziels meiner Masterarbeit und ein anderer Grund die Unbekanntheit der Methode und des Themas. Ersteres war problematisch, weil ich annahm, dass ich zum Erreichen meines Ziels, Conclusions-Abschnitte zu beurteilen, nur zwei Optionen hätte: Entweder direkt solche Abschnitte und deren Schluss-

folgerungen analysieren und dann ein Codebuch erstellen oder, umgekehrt, erst ein Codebuch erstellen konnte und dabei mehr über die Qualität lernen könnte. Da ich kein ausreichendes Verständnis von Schlussfolgerungen hatte, entschied ich mich dazu, mich auf die Erstellung des Codebuchs zu konzentrieren. Damit priorisierte ich das Codebuch vor dem Zweck dessens ein, mehr über die Qualität von Conclusions-Abschnitten herauszufinden. Dies hat dazu geführt, dass ich unermüdlich versucht habe, die inhaltlichen Strukturcodes zu verbessern und zu verfeinern, anstatt ungewöhnliche Stellen im Text aufzuspüren und genauer zu analysieren. Ich erkläre mir meine Entscheidung damit, dass das Thema unbekannt war und ich daher erst ein sehr gutes Verständnis über Conclusions-Abschnitte und Schlussfolgerungen erlangen wollte, bevor ich Aussagen über diese treffen wollte. Ich entschied mich also dazu, weitere Artikel zu codieren, bis ich genug gesehen hatte, um mir die Erkenntnisse darüber selbst zu erschließen – quasi durch Learning-by-Seeing. In diesem Prozess codierte ich und verglich dann die Codes. Gelegentlich hielt ich Ideen schriftlich fest. Dadurch wurden meine Tagesziele sehr repetitiv und fühlten sich nicht wie abgeschlossene Schritte an. Es war mir unklar, wie viele Schlussfolgerungen ich wohl anschauen müsste, um ein Muster darin zu erkennen.

Im Rückblick hätte ich Microcoding an interessant erscheinenden Stellen durchführen sollen, um mich gezielt auf die Erreichung des Ziels, mehr über die Qualität zu erfahren, vorarbeiten zu können. Außerdem hätte ich in regelmäßigen Abständen eine Art zusammenfassende Rückschau machen sollen, in der ich alle Vermutungen und Ideen in Bezug auf das Hauptziel hätte sammeln sollen. Dies hätte den Effekt gehabt, dass ich immer wieder an mein eigentliches Ziel erinnert worden, weniger abgeschweift wäre. Ich hätte das Gefühl gehabt, vorwärts zu kommen. Aus diesen Gründen wäre eine systematischer Anwendung von Microcoding hilfreich gewesen.

3.3.2 Suche nach Arten von Conclusions-Abschnitten

Nach dem initialen Lernen und Codieren sowie der Betrachtung des Codebuchs der qabstracts-Studie habe ich versucht, Muster innerhalb von Conclusions-Abschnitten zu finden. Das Codebuch sollte durch die Muster ergänzt werden und – ähnlich wie das Codebuch der qabstracts-Studie – damit Codes in eine Reihenfolge setzen und das Codieren vereinfachen. Die Muster sollten den inhaltlichen Fluss des Textes beschreiben. Weil Conclusions-Abschnitte generell weniger Strukturvorgaben haben als z.B. Abstracts, erwartete ich mehrere verschiedene Arten von Texten zu finden.

Dafür habe ich codierte Artikel betrachtet und versucht, Gemeinsamkeiten in der codierten Struktur zu erkennen. Innerhalb MAXQDAs konnte ich nur eine Tendenz von Struktur erahnen. Mir fehlte die Möglichkeit, mehrere Artikel nebeneinander betrachten zu können, weshalb ich alle bis zu dem Zeitpunkt codierten Artikel ausdruckte, um sie physisch nebeneinander legen und anordnen zu können. Am Ende entstand eine graduelle Sortierung nach den Längen und dem Vorhandensein bestimmter gleicher struktureller Inhalte, also strukturell inhaltliche Codes, die zu einer gemeinsamen übergreifenden Textrolle gezählt werden können (später Rollen-Codes, siehe 3.4). Beispielsweise würden zur übergreifenden Rolle einer Einleitung die Codes background, need und gap gehören. background bezeichnet Sätze, die Hintergrundinformationen geben, die bereits vor der Studie bekannt waren, need beschreibt

die Notwendigkeit der Studie und gap eine Lücke in der Wissenschaft, die vor der Studie bestand.

Nach intensiver Analyse der Conclusions-Abschnitte wurde sichtbar, dass Conclusions-Abschnitte verwandt zu Abstracts aufgebaut sind. Sie folgen grob demselben Textfluss mit wenigen Abweichungen. Die Abschnitte unterscheiden sich vor allem durch die Länge und den abweichenden strukturellen Inhalt. Das bedeutet, dass in Abstracts jeder Textteil bzw. jede Rolle immer vorhanden sein muss und alle Textteile eine ungefähr ähnliche Länge haben. Bei Conclusions-Abschnitten sind Textteile variabel in der Länge, je nach Geschmack des Autors, und beginnen nicht immer mit einer Einleitung. Das Ergebnis war überraschend, daher versuchte ich es mit neuen Artikeln zu wiederholen oder zu widerlegen. Ich stellte fest, dass der grobe Fluss von Conclusions-Abschnitten auch bei den neuen Artikeln ähnlich dem eines Abstracts war.

Textteile bzw. Rollen sind immer mindestens einen Satz lang, erstrecken sich i.d.R. aber über mehrere Sätze. Solche Rollen sind z.B. eine Einleitung oder ein Ausblick. Anfangs ging ich davon aus, dass ein Textteil immer genau einem Absatz entspricht. Es erschien mir logisch, dass Autoren Themen bzw. Rollen mithilfe eines Absatzes bündeln. Denn der Sprung zwischen zwei Absätzen spiegelt beim Lesen in der Regel einen Gedankensprung. Trotz dieser Eigenschaft beschränken sich Wechsel zwischen Rollen nicht auf Absätze, sondern können auch innerhalb dieser vorkommen.

Zusätzlich zu den Rollen habe ich versucht, die mentale Einstellung von Autoren zu erkennen oder zumindest festzustellen, welche Artikel ähnliche Einstellungen ausdrücken könnten. Mit mentalen Einstellung bezeichne ich Intentionen und Charaktereigenschaften der Autoren, wie z.B. die Bereitschaft, Schlussfolgerungen zu treffen oder welche Erwartungen sie an ihre Leser stellen. Codes zu finden die auf mentale Einstellungen hinweisen, gestaltete sich als schwierig. Meine ersten Ideen beschränkten sich auf die Erwartungen der Autoren an ihre Leserschaft bezüglich des Leseflusses. Mithilfe eines Codes, der die Ähnlichkeit eines Conclusions-Abschnitts zum Abstract beschreibt (likeabstract), und eines Codes, der die Abkürzungen kennzeichnet und angibt, ob diese erklärt werden oder nicht (abbrev), lässt sich ableiten, ob die Autoren erwarten, dass die Leser das Abstract vor den Conclusions lesen oder nicht.

Zusammenfassung Ich habe keine Muster gefunden, allerdings konnte ich feststellen, dass Conclusions-Abschnitte erstaunlich ähnlich strukturiert sind wie Abstracts. Conclusions-Abschnitte können mit einer Beschreibung zur Motivation der Studie und einer Beschreibung der Studie beginnen. Meist wird das Ziel der Studie zu Beginn der Studienzusammenfassung benannt. Nachfolgend werden Ergebnisse der Studie präsentiert. Dies ist der Teil in dem Schlussfolgerungen präsentiert und teilweise begründet werden. Zuletzt folgt in den meisten Fällen der Blick in die Zukunft. Eine durch Prof. Dr. Prechelt ausformulierte Beschreibung dieser Abfolge ist im Codebuch (siehe 3.4) zu finden und wurde in Rollen-Codes umgewandelt.

Auf der Suche nach Mustern habe ich bereits codierte Abschnitte verwendet. Alle meine Codes beinhalteten Aspekte der Struktur. Das könnte meine Ergebnisse beeinflusst haben, da ich dadurch stark auf die Struktur des Inhalts fokussiert war. Hätte ich schon Codes gehabt, die vielleicht nach Intentionen oder Selbstbewusstsein der

Autoren codiert wären, wäre ich möglicherweise in der Lage gewesen, Muster für mentale Modelle von Autoren erkennen zu können.

Mit der gefundenen Abfolge habe ich eine erste Version des Codebuchs erstellt, stark angelehnt an das Codebuch der qabstracts-Studie.

Selbstreflexion In dieser Phase fehlte mir eine Übersicht meiner Arbeit. Ich hatte viele Gedankenansätze, konnte sie jedoch nicht gut greifen. Mir schien es, als ob mir ein Austauschpartner fehlte, mit dem ich an diesem Projekt arbeitete und mit dem ich z.B. täglich zehn Minuten meinen Arbeitsfortschritt erfassen könnte. Da ich dafür leider keine geeignete Person zur Verfügung hatte, begann ich damit, ein ausführlicheres Forschungstagebuch zu führen. Die Option einer Entwickler-Gummiente, für die ich meine Gedanken ausformulieren hätte können, war aus persönlichen Gründen nicht möglich.

Möglicherweise habe ich mir bei der weiteren Theorienfindung bzgl. mentalen Modellen selbst im Weg gestanden, indem ich nur den sehr offensichtlichen Hypothesen nachgegangen bin und eher fragliche Hypothesen oft von vornherein verworfen habe.

Zu diesem Zeitpunkt hatte ich nur strukturelle Codes entwickelt, welche noch nicht detailliert genug ausgearbeitet waren. Einige Codes überschnitten sich, bei anderen waren die Grenzen noch unklar. Ein Beispiel dafür ist, dass ich zu diesem Zeitpunkt sicher war, dass es eine bestimmte, sich wiederholende Abfolge (result, conclusion und fxxx) innerhalb der Schlussfolgerungen gab. Dies schien mir, neben der zu Abstracts ähnlichen Struktur, die zweithäufigste Struktur innerhalb von Conclusions-Abschnitten zu sein. Mit der Endversion des Codebuchs und der erneuten Überprüfung meiner Codes fielen einige dieser codierten Abfolgen weg und stellten sich nicht als häufig wiederauffindbares Muster heraus.

3.4 Ergebnisse: Das Codebuch

Das nachfolgend präsentierte Codebuch ist die aktuellste Version. Das bedeutet, es ist mehr Arbeit hineingeflossen, als bisher beschrieben wurde. Daher möchte ich die Schritte zur Entstehung dieser Version kurz erläutern: Die ersten Versionen des Codebuchs entstanden zu Beginn meiner Masterarbeit und sollten mir helfen, mich möglichst gut innerhalb eines Conclusions-Abschnitts zurecht zu finden. Nachdem dies halbwegs gegeben war, versuchte ich Muster und Struktur in Conclusions-Abschnitten zu erkennen. Mit diesen Ergebnissen habe ich eine erste Version des Codebuchs erstellt. Als Vorbild diente das Codebuch der qabstracts-Studie. Die nächste Version des Codebuchs entstand als Vorbereitung des Studierendenprojekts und unterschied sich von vorhergehenden Versionen darin, dass ich die Komplexität der Anwendung des Codebuchs reduzierte, indem ich möglichst viele objektive Codes hatte und zum Erkennen dieser Codes Details und Beispiele zur Verfügung stellte. Die finale Version des Codebuchs für die Studierenden wurde von Prof. Dr. Prechelt erstellt. Im Anschluss wurde das Codebuch von den Studierenden selber noch verbessert und dann getestet. Dabei gefundene Mängel habe ich teilweise im Nachgang noch verbessert.

Die hier präsentierte Version enthält nur Codedefinitionen und inhaltliche Hinweise zum Codieren, keine projektspezifischen Hinweise zu syntaktischen Codierungs-

feinheiten. Das Codebuch wurde auf Englisch verfasst, um später für einen größeren Forscherkreis zugänglich zu sein.

Rules for the codingprocess

1. Simplicity: We want a coding system that can be applied rapidly and reliably. To achieve this, we will often prefer simplicity over precision and richness.
2. Granularity: Annotation granularity is by sentence. Sentences are terminated by a colon or a period.
3. An annotation usually consists of two codes (separated by a comma): A role code and a content code (see below). One or more optional codes (starting with a dash, see below) may follow.
4. If necessary, a sentence (in particular a long one) can receive two or more content codes. This means that some substantial part(s) of the sentence pertain to one code and other substantial parts pertain to another code. An interpretation with only one code is prefered. No part of any sentence pertains to two content codes at once.
5. For annotation, a sentence is interpreted in its backward and 1-sentence forward context, not in isolation. In particular, if a subsequent sentence provides detail that could have been part of the previous one, both will often get the same code.
6. In particular, when a sentence can be an A or a B, but only a B is expected in this position, it is interpreted as a B.
7. We are gentle in detecting negative aspects: We annotate them only where they are clearly present.
8. The role codes and content codes are meant to be (nearly) objective; we strive for high inter-coder agreement. In contrast, some of the optional codes and suffixes are subjective.

Role-Codes (dt.: Rollen-Codes)

There is a typical sequence of concerns that are treated in a Conclusion section, although not all are present in every case: Orientation, Study Summary, Outcome, Outlook. Each sentence contributes to exactly one of these, as indicated by the role code applied to the sentence. Main roles:

- **code OR: Orientation.** This sentence explains the study's topic area or motivation.
- **code SS: Study Summary.** This sentence summarizes some of the study's goals, methods, or immediate results.
- **code OC: Outcome.** This sentence points out or interprets what the present study has found.

- **code OL: Outlook.** This sentence talks about the future as it is influenced by the present study.

We define additional auxiliary role codes for the purpose of executing our study: **code TC: Table of Contents**, used for the list of headings at the top of the extracted file and **code OT: Other**, used for material that does not belong to any of the other roles, such as captions of figures or tables that pertain to other sections than the Conclusion section.

Content-Codes (dt.: Inhalt-Codes)

Below, the content codes are listed under only one role, but that is only their typical position. Be prepared to find such sentences in other places as well, which will often (but not always) indicate a quality problem.

Content-Codes for Orientation

- **code background**: Information about the larger topic area of the work, the state-of-the-art, definitions, related work and what is known.
- **code gap**: A statement of what is not yet known or what is so far difficult to achieve in research or in practice which directly leads over to the topic of the present work (the objective).
- **code need**: A postulate about what research needs to be performed which directly leads over to the topic of the present work (the 'objective').

Content-Codes for Study Summary (SS)

- **code objective:abbrev**: One or more of this work's top-level research goals, specific top-level research interests, or top-level research questions. Objectives sentences often contain material that is ambiguous as goal-or-method. In those cases, assign two codes only if there is no method information elsewhere in the Study Summary. If the method information is too specific and concrete to be a goal, assign two codes in any case. If a study does a systematic literature research, just doing such a research already counts as objective. The work towards that objective has now been done, so the formulation can be as an objective, past objective, or reached objective. There is usually an objective sentence at the beginning of a Study Summary, but more of them may follow later.
- **code method:abbrev:comparison:defensive**: Information about the approach or setup of the empirical (or possibly purely mathematical) study, what were the steps taken to reach the objective. This also includes artifacts, including software, built to execute the study. Such artifacts should not to be coded as -artifact.
- **code result:abbrev:comparison:defensive**: Information about the immediate outcome of a study in the form of empirical results. See optional code -teaser.

Content-Codes for Outcome (OC) Common codes in this stretch are result and less often also fposs and fshould.

- code conclusion-ex:comparison:defensive:high:low: A take-home message that is less specific than one or more results and that takes the form of an existence proof statement that was not obvious before the present study. If the statement appears to be an existence statement only because the authors shy away from formulating a generalization, but *should* rather provide a generalization, code it as conclusion-gen:low:defensive instead.
- code conclusion-gen:comparison:defensive:high:low: A take-home message that is less specific than one or more results and that takes the form of a generalization from the results. To be a generalization, a conclusion must be written in present tense, not past tense, nor must it be constrained to the specific case investigated in the study. Both of these would make the same sentence a result summary.
- code limitation:defensive: Information about limitations, threats to validity, and the like of the study methods or its results, including factors that could affect the reliability or generalizability of the study. Use :defensive' when the statement is made to avoid drawing a (stronger) conclusion.
- code artifact: An artifact is a product of the article which can be used or understood without the article (i.e. guide, taxonomy, method, software). Only if the sentence does nothing else than describing an artifact (i.e. how it was built or made) we code it as artifact. If the sentence describes an artifact and its impact, code it accordingly to the described impact (for „With this artifact things could happen.“ fposs.) See optional code -teaser.

Outcome may be difficult to delineate, because results bleed into it around conclusions, methods bleed into it around artifacts, conclusions also occur in Outlook, results and conclusions can be hard to discriminate, and generally stray sentences of almost any kind sometimes occur in curious places in a Conclusion section.

Content-Codes for Outlook (OL) fxxx sentences have taken one step further than conclusions, they always contextually emerge from a study learning. Every fxxx sentence describes some kind of intention, while a conclusion-gen is either a rule or description of a state.

- **code fposs:defensive**: A neutral statement about what future research (incl. development, understanding) is now possible (i.e., could now be done, by whoever). Note: Statements regarding what *practitioners* can now do are conclusion!
- **code conclusiontail**: A statement about a rather obvious effect of possible study outcomes. May appear similar to conclusion or fposs on first glance, but do not deliver new information regarding the objective or some future directions.
- **code fhope**: A statement where the authors express the hope that somebody will do a certain future research.

- **code fshould**: Statement about what future research should be done (by whoever).
- **code fwill**: A statement where the authors express the convicition that something will happen in the future.
- **code fwork**: Explicit statement about what future research (or research-related tasks) the authors intend to do.

generic Content-Codes These can be found with any of the roles.

- **code missingcode**: To be used in the open coding phase only, not the content analysis phase. The sentence contains information that cannot be represented adequately by the existing codes: We should extend the codebook to allow for capturing this information.
- **code notcoded**: This sentence has been intentionally ignored.
- **code relatedwork**: An entire sentence about studies besides the current one. Use suffix :comparison when other studies are related to the present one.
- **code resourcepointer**: A reference to a concrete external resource such as a software artifact, materials package, data package, appendix or similar item.
- **code transition**: A sentence that only serves to finish or lead over to a different thought without providing further information.
- **code summary**: A statement that summarizes several results, but does not provide new information. A summary statement does not generalize beyond the immediate results.
- **code badgrammar**: A sentence that is harder to understand because of its incorrect grammar. Remember to code gently, especially when using subjective codes on sentences with odd grammar.

Suffix

- **objective suffix :abbrev** : The sentence uses an abbreviation that is specific to the present study and it (or the next sentence) explains the meaning of the abbreviation.
- **objective suffix :comparison** : The sentence compares an element (such as a result) of this study to one or more specific previous studies.
- **suffix :defensive**: The sentence plays down a result, the capabilities of a method, the possibilities for future research or a conclusion. Sometimes the topic that is meant is not in the immediate sentence.

- **suffix :high and suffix :low**: To be used when a conclusion is expressed such that it lacks confidence (:low) or such that it exhibits unusually or inappropriately high (:high) confidence. Use no suffix if neither is the case, which should be the most common case.

Optional Codes These codes can never occur alone, only in conjunction with one of the others. They need not match the fellow coders coding. They start with a dash to signal their special status and should be given last in any set of annotation braces. Like every other code they should be seperated by a comma (i.e. conclusion, -goodone).

- **code -likeabstract**: If a sentence is very similar (or even the same) two a sentence in the abstract.
- **code -teaser**: The sentence hints at information, missing in the conclusion-section, in a manner that may pique the reader's curiosity, e.g. „We have found 7 ways how X influences Y.", „As explained in Section 4.2, ..."
- **code -goodone and code -badone**: Mainly meant ot be used on conclusions to mark good conclusions and bad conclusions, but can also be used for other good/bad samples. These codes should not be included in the content analysis. They are meant to help collect good and bad samples to help us learn more about those.

4 Erprobung des Codebuchs durch Studierende

Nach Beginn meiner Vorbereitungsphase schlug mein Betreuer mir vor, an einem Teil seines Tutoriums zum Kurs „Empirie im Softwareengineering" mitzuwirken. Der Kurs vermittelt Studierenden der Informatik einen Überblick über geläufige empirische Methoden und bietet die Möglichkeit, eine dieser Methoden anzuwenden. In dem damals bevorstehenden Sommersemester 2023 war geplant, die Studierenden eine Inhaltsanalyse ausprobieren zu lassen mit offenem Codieren und einer anschließenden quantitativen Phase, in welcher viele Conclusions-Abschnitte mit den zuvor erstellten Codes codiert und dann Analysiert werden sollten.

Da offenes Codieren ein zeitaufwändiges Verfahren ist, sollten die Studierenden nicht von Grund auf neu beginnen müssen, sondern mit dem von mir bereits begonnen Codebuch arbeiten können. Die Idee war also, dass die Studierenden eine vereinfachte Phase des offenen Codierens bekommen sollten und ich nach dem Kurs möglichst viele codierte Artikel, die ich analysieren könne. Dafür wäre ein fast fertiges Codebuch mit Codes, die die Qualität beschreiben könnten, optimal gewesen. Meine Vorbereitung bestand darin, dass ich das Codebuch überarbeitete, eine Einführung gab und den Beginn der Forschungsphase begleitete. Ein Arbeitsablauf fürs Codieren in der Gruppe und eine dazu passende digitale Codierungs-Umgebung wurden von Prof. Dr. Prechelt gestellt. Es war nötig einen Arbeitsablauf einzuführen, da der spätere quantitative Codierungsprozess so ablaufen sollte, dass jeder Abschnitt von zwei separaten Studierendengruppen codiert und anschließend abgegelichen werden

sollte. Außerdem sollte es möglich sein, dass sich alle elf Studierenden dabei so absprechen konnten, dass sie z.B. nicht alle versehentlich denselben Artikel codieren würden. Damit wurde ein Arbeitsablauf und eine Umgebung benötigt in der sich koordiniert werden konnte und in der Conclusions-Abschnitte doppelt codiert werden konnten. Die Umgebung, die Prof. Dr. Prechelt eingeführt hat, hat dies gelöst, indem sie Conclusions-Abschnitte extrahiert und daraus jeweils zwei Dokumente erstellte, innerhalb derer in vorgegebenen Stellen codiert werden konnte. Um Artikel zu codieren mussten die Studierenden weder die Artikel importieren, noch den Conclusions-Abschnitt suchen oder die Überschriften, die zu betrachten waren. Wer welche Artikel codierte und mit wem abglich, war mit wenig Aufwand koordiniert. Jeder zu codierende Abschnitt enthielt durch die Art der Darstellung Zeilennummern, die das abgleichen per Hand stark vereinfachten. Die Codierungsumgebung enthielt außerdem ein Skript, welches Zugriff auf das Codebuch hatte und alle vergebenen Codes mit den dort beschriebenen Codes auf deren Existenz verglich, so dass eine Datenbereinigung bezüglich falscher Benennungen vor Analyse der codierten Artikel minimal gehalten wurde. Zusätzlich gab es ein Skript, welches die vergebenen Codes zwischen zwei unterschiedlichen Codierergruppen verglich, um den Codeabgleich in der quantitativen Codierungsphase zu vereinfachen.

Sowohl der Arbeitsablauf, als auch die damit verbundene Umgebung, hatte Prof. Dr. Prechelt bereits für seine qabstracts-Studie entworfen. Die Aufgabe, die Umgebung von Abstracts auf Conclusions anzupassen, hat er übernommen.

4.1 Forschungsarbeitsbericht über das Studierendenprojekt

Die Beschreibung der Umsetzung der Zusammenarbeit mit den Studierenden erfolgt chronologisch in diesem Abschnitt. Es wird jeweils mit dem Forschungsbericht begonnen wird. Dann folgt eine kurze Zusammenfassung von Zwischenergebnissen, die in den meisten Fällen auf einen anderen Abschnitt verweist. Die Ergebnisse und Erkenntnisse dieser Phase sind in Abschnitt 4.3 ausführlich beschrieben.

4.1.1 Vorarbeiten zum Kurs

Vor Beginn des Projekts habe ich mehrere Maßnahmen ergriffen, um die Arbeit der Studierenden zu erleichtern. Ich habe die Komplexcität meines Codebuchs reduziert. Des Weiteren habe ich die ersten Tutorien zu Beginn des Projektes mitgeplant und mir dabei Gedanken darüber gemacht, wie man den Studierenden das Codieren erleichtern könnte.

Vereinfachungen am Codebuch: Das Codebuch sollte ein Snapshot aller meiner Ideen und Gedanken werden, sodass Studierenden an den aktuellsten Ideen weiterarbeiten könnten. Das bedeutet, dass das Codebuch sehr umfangreich war, da ich Hinweise und Vermutungen zu Codes beschrieb. Die Hinweise basierten auf meiner Codiererfahrung und standen deshalb im Einklang mit den bis zu dem Zeitpunkt von mir codierten Artikeln. Nicht überprüfte Hinweise waren deutlich als Vermutungen gekennzeichnet.

Für die Vereinfachung des Codebuchs habe ich versucht, so viele Codes wie möglich objektiv zu halten und möglichst wenige subjektive Codes zu inkludieren oder Codes, die zusätzlich zu anderen Codes genutzt werden, wie Suffixe oder optionale Codes. Um dies zu erreichen, habe ich mich auf die strukturell inhaltlichen Codes fokussiert. Strukturell inhaltliche Codes sind objektiv und haben den Anspruch unabhängig von subjektiven Einschätzungen Codierender zu sein. Die meisten meiner subjektiven Codes erschienen mir zu spekulativ, weshalb ich sie nicht ins Codebuch aufnahm. Manche der aufgenommenen subjektiven Codes wurden von Prof. Dr. Prechelt in objektivere Codes umgebaut. Bspw. wurde der Code justification, der einen Satz beschreibt, in dem Autoren ihre eigene Arbeit in besseres Licht rückten, zum Code comparison, der einen Vergleich signalisierte, der gemacht wurde. Dies ist für viele der Stellen, an denen justifaction vergeben wurde, ausreichend, da dieses Hervorheben der eigenen Arbeit oft durch einen Vergleich erzeugt wird.

Weiterhin habe ich Codes zusammengefasst, deren Unterscheidung kaum einen Mehrwert bringt. Z.B. habe ich den Code transition eingeführt, welcher zuvor aufgeteilt war in introduction sentence und out-troduction sentence. Außerdem habe ich Hilfscodes entfernt und diese in Hinweise beim codieren umformuliert: So war der Hilfscode Explanantion of following or previous durch die Codierungsregel 5 (siehe 3.4) im Codebuch gedeckt, dass immer mit Beachtung des Kontexts codiert werden soll.

Das so entstandene Codebuch testete ich, vor dem Tutorium, indem ich es auf weitere Artikel anwendete. Dabei achtete ich vor allem auf nicht leicht codierbare Sätze und missverständliche oder ungenaue Codedefinitionen. Dies zeigte mir einige Stellen im Codebuch, die noch nicht ausgereift waren. Einige davon konnte ich verbessern, andere markierte ich als unfertig und damit als Aufgabe für die Studierenden beim offenen Codieren. Beispielsweise ist mir hier die unzureichende Definition des Begriffs objective aufgefallen, die noch nicht berücksichtigte, dass ein Ziel in den Schlussfolgerungen auch in der Vergangenheitsform formuliert werden kann, da es am Ende der Studie bereits erreicht sein kann.

Vorbereitungen: Zur Vorbereitung auf den Kurs und die Begleitung der Studierenden habe ich mich mit dem qabstracts-Repository vertraut gemacht. Beim gemeinsamen Codieren konnte ich die Anwendung der Codes durch die Studierenden überprüfen und ggf. gemeinsam mit ihnen korrigieren. Auf diese Weise sollte sichergestellt werden, dass das Codebuch schnell erlernt werden kann.

Als letzten Vorbereitungsschritt erstellte ich eine Präsentation, um den Studierenden das Forschungsziel meiner Masterarbeit und ihren Anteil daran vorzustellen. Zusätzlich enthielt die Präsentation in ihrer ursprünglichen Form auch Erläuterungen zum Codebuch und eine kurze Vorstellung dessen.

Kurz vor Kursbeginn fand Prof. Dr. Prechelt noch Zeit, das Codebuch durchzusehen und zu überarbeiten. Aus meiner ausführlichen Momentaufnahme des aktuellen Forschungsstandes wurde ein kurzes, stark strukturiertes und wesentlich weniger komplexes Codebuch. Meine Version war so gestaltet, dass Codes ausführlich definiert und Vermutungen ausformuliert waren, so dass sie leicht erweiterbar waren. Vermutungen, Hinweise und Definitionen wurden so erläutert, dass der Kontext leicht ersichtlich war. Prof. Dr. Prechelts Version hatte dagegen die Definitionen und

Hinweise verkürzt und Vermutungen aus dem Hauptteil an den Schluss des Codebuchs verschoben. Vermutungen waren zu Aufgaben umformuliert worden, die keine offenen Gedanken enthielten, und wurden in einem eigenen Abschnitt zusammengefasst. Dadurch war das Codebuch wesentlich übersichtlicher.

Der Gedanke, den Prof. Dr. Prechelt dabei verfolgte, war, dass vorgegebene Vermutungen und Gedanken zu sehr eine Richtung vorgaben und die Studierenden beeinflussen würden. Studierende, die noch nie zuvor codiert hatten, würden mit vorgegebenen Vermutungen in eine Richtung gelenkt. Damit wäre keine uneingeschränkte Ideenentwicklung mehr möglich, was dem offenen Codieren widerspricht. Mit der neuen Version ist eine uneingeschränkte Ideenentwicklung möglich. Außerdem konnte Prof. Dr. Prechelt vereinzelte Codegruppen effizienter gestalten.

Die Aufgaben für die Studierenden in ihrer offenen Codierungsphase sind im Anhang B zu finden.

4.1.2 Arbeiten während des Kurses

Insgesamt war für den Forschungsteil des wöchentlichen Tutoriums fast das gesamte Sommersemester vorgesehen. Elf Studierende nahmen am Projekt teil. Ich war vor allem in der Vorbereitung des Forschungsteils eingebunden und in den ersten zwei Wochen des Kurses. Innerhalb der ersten zwei Wochen sollte die offene Codierungsphase stattfinden und das Codebuch abgeschlossen werden. Ich half bei der Planung und Durchführung von zwei Tutorien und bei der generellen Organisation, sowie dem Codieren. In jedem Tutorium habe ich den Studierenden die Möglichkeit gegeben die Vorwoche zu besprechen. Ich habe versucht die Studierenden zur Mitarbeit anzuregen, indem ich Stimmungsbilder mit einfachen Handzeichen machen ließ. Für Retrospektiven aus Scrum [86] sagt man, dass ein solches einfaches Meinungsbild am Anfang wie ein Eisbrecher wirkt [16]. Im Anschluss an die offene Codierungsphase setzten die Studierenden die Codierung und Analyse alleine fort. Ab diesem Zeitpunkt führte der Tutor den Kurs wieder alleine.

Einführung in das Studierendenprojekt: Im ersten Tutorium habe ich mit Unterstützung von Prof. Dr. Prechelt eine Einführung in das Thema gegeben. Ich habe das Ziel meiner Masterarbeit und den Beitrag der Studierenden erklärt. Lernziel der Studierenden war es, die Anwendung von Techniken der GTM zu erlernen, wie offenem codieren und theoretischer Sensitivität. Der Beitrag zu meiner Arbeit war das Codebuch zu verbessern und zu testen, indem sie quantitative Daten mit dem Codebuch sammelten und anschließend analysierten.

Offene Codierungsphase des Studierendenprojekts: In der ersten Woche hatten die Studierenden die Aufgabe, ca. vier Conclusions-Abschnitte zu codieren, dabei das Codebuch kennenzulernen und bereits erste Codebuchverbesserungen vorzuschlagen, die im folgenden Tutorium besprochen werden sollten. Meine Aufgabe dabei war es als zweite codierende Person, mit Gruppen von zwei bis drei Studierenden zu arbeiten und gemeinsam codierte Artikel abzugleichen. Mit jeder Gruppe habe ich drei bis vier Conclusions-Abschnitte verglichen und Verständnisse der Codes ausgetauscht. Je nach Gruppe war es mir möglich während dieses Austauschs mehr über

den Umgang meiner Codes zu lernen und den Studierenden ein besseres Verständnis der Codes zu vermitteln. Dadurch konnten Formulierungen von Codedefinitionen unmissverständlicher gestaltet werden. Dies funktionierte gut, wenn das Abgleichen unterschiedlicher Codes Diskussionen auslöste, welche den Kontext des Textes miteinbezogen. Weniger gut funktionierte dies wenn der Codeabgleich eher mechanisch mit Hilfe eines Skripts passierte und Studierende aktiv Diskussionen vermieden, indem sie z.B. eigene Begründungen nicht nannten und direkt meiner Codierung zustimmten. Dadurch war es kaum möglich solchen Studierenden ein besseres Codeverständnis mitzugeben.

Ich habe gemeinsam mit dem Tutor des Kurses das folgende Tutorium vorbereitet. Wir haben den Ablauf geplant und Online-Tools für den Austausch zwischen den Studierenden sowie weitere Arbeiten an dem gemeinsamen Codebuch eingerichtet. Die Möglichkeit zu Codieren und die Codes abzugleichen wurde, bereits von Prof. Dr. Prechelt bereitgestellt. Wir haben bei der Planung des Tutoriums die Struktur einer Scrum-Retrospektive [86] verwendet. Das bedeutet, wir haben erneut mit einem Check-In, in Form eines Wochenrückblicks, begonnen und ließen eine Diskussion folgen.

Der wöchentliche Rückblick zeigte, dass das Verständnis des Codebuchs durch die Vorarbeit mit der qabstracts-Studie erleichtert wurde und auch Codes mit ähnlichem Namen, aber nur leicht unterschiedlicher Definition kein Problem darstellten, da die Studierenden bereits im Voraus darauf hingewiesen wurden. Mehrere Studierende berichteten, dass ihnen aufgefallen war, wie sie beim Codieren Rollen-Codes entsprechend der Zugehörigkeit der Inhalt-Codes vergeben haben, ohne darüber nachzudenken, ob der Rollen-Code an dieser Stelle passend war. Dies könnte in der quantitativen Phase zu falsch vergebenen Rollen-Codes geführt haben.

Nach einer ein-wöchigen Lernphase für das Codebuch sollten die Studierenden Verbesserungen am Codebuch vorschlagen. Diese Lernphase umfasste das Codieren mithilfe des von mir gestellten Codebuchs und das Abgleichen der codierten Artikel mit mir, um mehr über die Codes zu lernen. Zusätzlich dazu hatten die Studierenden die Hausaufgabe, Ansätze zur Verbesserung von Codes zu erarbeiten. Drei dieser Ansätze sollten gemeinsam im Tutorium diskutiert werden. Da es mehr als drei solcher Verbesserungsansätze gab, mussten drei Ansätze ausgewählt werden. Dies wurde per Abstimmung getan. Aufgrund von Prof. Dr. Prechelts Einwand wurde eines der meist gevoteten Themen exkludiert, welches die Zusammenführung der Codes fhope und fshould betraf. Stattdessen wurde eines der anderen Themen, über das am meisten abgestimmt wurde, zwei Mal diskutiert. Die insgesamt elf Studierenden wurden in drei Diskussions-Gruppen aufgeteilt. Jede Gruppe hatte 30 Minuten Zeit. Die Diskussionen waren vielversprechend.

Das doppelt besprochene Thema war eine neue Dimension für den Code conclusion. Mit der Dimension sollte die Bestimmtheit der Schlussfolgerung festgehalten werden können. Gruppe A hat versucht, Signalwörter zu finden, die eine niedrige oder hohe Bestimmtheit andeuten, während Gruppe B überlegt hat, in wie viele Stufen die Bestimmtheit einer Schlussfolgerungen eingeteilt werden kann. Die dritte Gruppe behandelte das Thema des Codes limitation. Es wurde erörtert, was Einschränkungen sind und welche Arten es davon gibt. Die Gruppe arbeitete nah an den Daten anhand von Beispielen und kam zum Schluss, dass limitation für Ergebnisse, Schlussfolge-

rungen sowie Methoden existieren können. Keine der Gruppen konnte die Ideen innerhalb des Tutoriums so weit voranbringen, dass konkrete Codebuch-Veränderungen zustande kamen, aber alle Gruppen haben einleuchtende Ideen verfolgt und damit erfolgreich offenes Codieren angewandt. Zum Abschluss des Tutoriums wurden die einzelnen Gruppen im Plenum angehört und Ergebnisse diskutiert.

In der folgenden Woche hatten die Studierenden Zeit, gefundene Ansätze weiter zu diskutieren und in konkrete Verbesserungsvorschläge in Form einer konkreten Pull Request (PR) für das Codebuch umzuwandeln. Diskussionen und Abstimmungen, die den konkreten Verbesserungen vorausgingen, fanden online statt. Diese PRs sollten von mir als Qualitätsportal betrachtet und zusammengeführt oder begründet abgelehnt werden.

In dieser Woche habe ich außerdem die quantitative Codierungsphase vorbereitet, indem ich die von Prof. Dr. Prechelt bereits extrahierten Texte von Conclusions-Abschnitten bereinigte, sodass diese bereit zum Codieren waren. Ich bin bis zum damals aktuellen Stand der qabstracts-Studie gekommen. Damit waren etwa 100 Conclusions-Abschnitte bereit, codiert zu werden.

Von den Verbesserungsvorschlägen wurden etwa die Hälfte in konkrete PRs umgesetzt, während der Rest nur diskutiert, vorgeschlagen und abgestimmt wurde. Vorschläge, die keine PR hatten, aber als positiv abgestimmt waren, habe ich nach Absprache trotz fehlender PR ins Codebuch nachgetragen. Denn positiv abgestimmte Vorschläge waren ausformulierte relevante Codebuchverbesserungen für die kein Studierender die Zeit gefunden hatte, sie in eine PR zu schreiben. Die gestellten PRs waren alle sinnvoll. Es gab keine Vorschläge, bei denen der Aufwand in einem ungünstigen Verhältnis zum möglichen Nutzen stand oder bei denen sich die PRs widersprachen. Dies war wahrscheinlich eine Folge der Diskussion im Tutorium, in der solche Verbesserungsvorschläge erkannt und abgefangen wurden. Nur in einem Fall musste ich die Formulierung der Vorschläge verfeinern.

Damit hatten die Studierenden das offene Codieren abgeschlossen und das Codebuch vorerst fertiggestellt. Nun ging es in die quantitative Phase, an der ich nicht beteiligt war.

4.1.3 Nacharbeiten zum Studierendenprojekt

In der Nachbearbeitung begann ich damit, meine bereits codierten Artikel auf Basis des neuen Codebuchs erneut zu codieren. Anschließend las ich meine Memos durch.

Wenig überraschend konnte ich vor allem in den zu Beginn meiner Arbeit codierten Conclusions-Abschnitten einige falsch codierte Stellen finden und korrigieren. Ich erkläre mir das damit, dass mein Codebuch am Anfang meiner Forschung noch nicht ausgereift war und ich auch das Codieren noch nie zuvor angewandt hatte. Durch das gesamte Codieren sind mir einige unausgereifte Stellen (z.B. ungenaue Definitionen) im Codebuch aufgefallen, die ich verbesserte. Da ich eine gute Übersicht über alle meine codierten Artikel hatte, konnte ich Vermutungen aus den Memos direkt überprüfen. Durch das Lesen der Memos wandte ich mich intuitiv der Verbesserung des Codebuchs zu. Durch die nun dazu kommende Analyse der Studierenden über ihr Projekt bekam ich weitere konkrete Hinweise zu noch nicht ausgereiften Stellen im Codebuch. Dabei habe ich mich aufgrund der Beschaffenheit des Codebuchs auf die

strukturellen Codes konzentriert und übersehen, dass die strukturellen Codes wenig über die Qualität der Schlussfolgerungen aussagen.

Ich habe systematisch Hypothesen aus Memos mit meinen bis dahin codierten Artikeln verglichen und habe viele Hypothesen widerlegt, weshalb ich zu dem Schluss gekommen bin, dass meine strukturellen Codes keine Aussagekraft haben. Folgende Erkenntnisse wurden aus dem Überprüfen der Memos gewonnen:

- Verschiedene Wörter für denselben Begriff des Artikels (z.B. Artikel,Paper) stehen nicht im Zusammenhang mit der Qualität von Artikeln.
- Artikel, die einen schlechten Conclusions-Abschnitt haben, haben nicht zwingend auch ein schlechtes Abstract. Ein schlechtes Abstract deutet nicht auf einen schlechten Conclusions-Abschnitt hin.
- Meine zu Beginn in fast jedem zweiten Artikel erkannte Struktur in Conclusions-Abschnitten kommt deutlich seltener vor, als ursprünglich angenommen. Dies liegt daran, dass meine initial definierten Codes nicht genau genug waren. Die gesehene Struktur lässt sich mit folgender Abfolge beschreiben: result, conclusion, fxxx.
- Wir trennen laut Codebuch nach „:“ und „.“. Meine Vermutung, dass auch „;“ als Trennzeichen gesehen werden sollte, hat sich als falsch heraus gestellt. In allen Fällen meines Samples wird „;“ nur als eine Trennzeichen für Aufzeählungen verwendet. Es werden keine Sätze voneinander abgegrenzt.
- Die Chance, dass ein Conclusions-Abschnitt mit der Überschrift „Conclusion**s**“ qualitativ hochwertig ist, ist höher als bei einem Abschnitt mit der Überschrift „Conclusion“ ohne „s“.
- Abschnitte und Gliederungen mit ungewöhnlichem Wortlaut haben eine deutlich höhere Wahrscheinlichkeit auch entsprechend ihrer Beschreibung gefüllt zu sein. Gewöhnliche Abschnittsüberschriften sind „Conclusion(s) (+ future work)“ und Überschriften die „Conclude“ enthalten oder „show/summarize conclusion(s)“ lauten. Eine ungewöhnliche Überschriften ist z.B. „Learned Lessons and Future Work “

Nach der erneuten Codeverfeinerungen und dem kompakten Codieren vieler Artikel auf einmal, hatte ich nun ein Verständnis von Schlussfolgerungen dem ich genug vertraute, um es weiter zu nutzen und Vermutungen zu Schlussfolgerungen überhaupt erst zuzulassen. *Schlussfolgerungen* sind entweder Generalisierungen oder Existenzbeweise. Sie tragen immer etwas Essentielles zum Wissen des Lesers über das Thema bei und sind immer im Präsens formuliert. Häufig folgen sie aus benannten Resultaten. Sätze die nur bestimmte Effekte beschreiben, aber keine konkreten Informationen liefern, sind keine Schlussfolgerungen.

Gerne hätte ich eine Analyse der am häufigsten verwendeten Wörter innerhalb bestimmter Codierungen durchgeführt, z.B. bei den Schlussfolgerungen. Leider funktioniert das nicht, da es zu viele codierte Stellen gibt, die ihren Code nur aufgrund des Kontextes erhalten haben und in Isolation codiert einen anderen Code erhalten hätten. Deshalb sind nicht alle codierten conclusion Sätze direkt typische Ausprägungen des Codes conclusion und verfälschen eine Wortanalyse.

4.2 Ergebnisse des Studierendenprojekts – quantitative Analyse

Insgesamt haben die studierenden 99 verschiedene Conclusions-Abschnitte codiert. Zur Gewährleistung größtmöglicher Korrektheit wurde jeder Abschnitt von zwei separaten Gruppen codiert und anschließend deren Ergebnisse abgeglichen. Dies wurde mit anders zusammengesetzten Gruppen wiederholt.

Folgend werden die Ergebnisse der Studierenden besprochen. Diese sind innerhalb des Abschlussberichts der Studierenden zu finden. Der Bericht ist eine Zusammenfassung der Arbeit der Studierenden, den die Studierenden zum Abschluss ihres Projekts erstellt haben. Der komplette Bericht ist online zu finden [89]. Der folgende Abschnitt orientiert sich nach der Präsentation der Ergebnisse der Studierenden.

Sequenz Analyse: Zur Sequenz der Rollen-Codes fanden die Studierenden raus, dass die meist verwendeten Sequenzen in Conclusions-Abschnitten OR, SS, OC, OL oder SS, OC, OL sind und beide jeweils ein Viertel der gesamten analysierten Menge ausmachen. Insgesamt enthalten ca. 70 % der Artikel die Abfolge SS, OC. Dies erscheint einleuchtend, da Artikel so ihre Zielstellung wiederholen, um dann die Ergebnisse darauf folgen zu lassen. Diese Ergebnisse bestätigen meine Theorie, dass Conclusions-Abschnitte eine ähnliche Struktur zu Abstracts aufweisen. Der Unterschied liegt hauptsächlich in der Strenge der einzelnen Strukturteile.

Im Abstract besteht die Tendenz, dass jeder Strukturteil enthalten sein sollte, während im Conclusions-Abschnitt nicht alle im Abstract vorkommenden Teile relevant sind, auch wenn die Abfolge starr bleibt. Jedes Abstract sollte bspw. mit einer Einführung des Hintergrundes und der Motivation zur Arbeit beginnen. Im Conclusions-Abschnitt ist dies jedoch überflüssig. Folglich ist der OR-Teil optional und wird auch nur von 41 % aller analysierten Artikel genutzt. Im Gegensatz dazu werden Ziel und Methodik der Studie (SS), obwohl sie bereits im Abstract und im Körper der Arbeit genannt wurden, häufiger wiederholt und werden in ca. 70 % aller Sequenzen den Outcomes (OC) vorangestellt. Das Ziel nochmals zu benennen und anschließend auf die daraus gewonnen Erkenntnisse einzugehen, um dem Leser eine bessere Verständlichkeit zu ermöglichen, ist eine gängige Vorgehensweise. Die Sequenzanalyse, welche lediglich die Top-fünf Sequenzen präsentiert, umfasst etwa 75 % der insgesamt analysierten Daten. Die Sequenzen der restlichen Daten (25 %) weisen repetitive Strukturen auf, in denen sich zwei Rollen-Codes abwechseln. Dies deutet auf eine scheinbar seltener vorkommende Form von Conclusions-Abschnitten hin, die neben denen existieren, die ähnlich wie Abstracts aufgebaut sind. Informationen sind nicht nach Forschungsschritt, sondern nach Thema zusammengefasst. Der Text wird um eine Schlussfolgerung herum strukturiert, anstatt dass rein linear Ziele, Methodiken, Resultate und dann alle Schlussfolgerungen gesammelt abgehandelt werden. Allerdings lassen sich damit nicht alle der beschriebenen verbleibenden Sequenzen der Studierenden erklären. Sequenzen bei denen sich OC und OR abwechseln scheinen mir in der Nachbetrachtung und Nachprüfung zufälliger Artikel eher Fehlcodierungen zu sein, die durch zu wenig Kontextaufmerksamkeit entstanden sind. Der Code background, der typischerweise in OR gefunden wird, kann z.B eine Hintergrundinformation sein, die einen Begriff eines Schlussfolgerungssatzes erklärt. Ist dieser andere Satz eine conclusion, sollte der Satz mit Hintergrundinformationen auch als conclusion codiert werden, da dieser die Schlussfolgerung erweitert und ohne diese

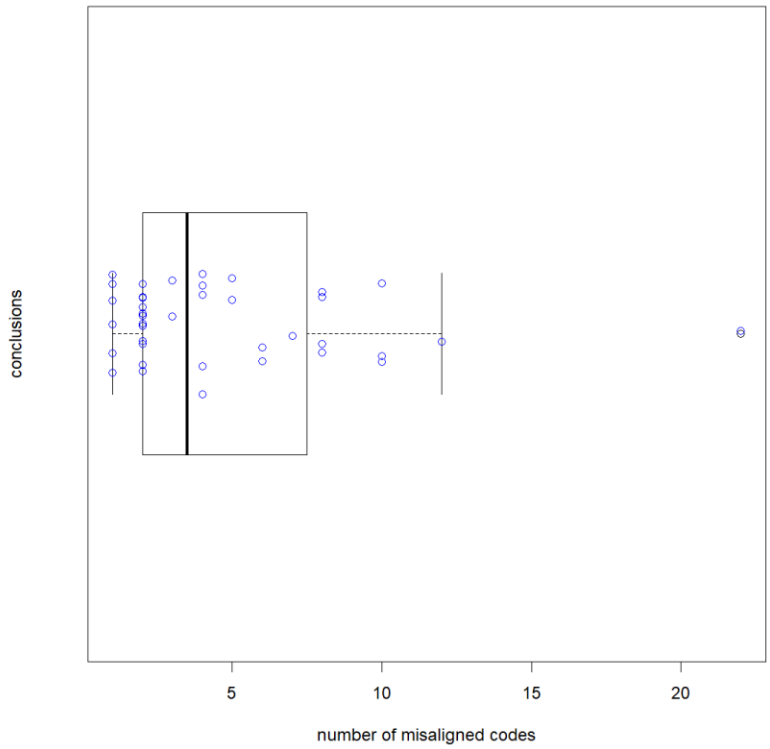

Abbildung 1: Abbildung der Verteilung von „misaligned role codes“ in Conclusions-Abschnitten nach Anzahl. Jeder Punkt entspricht einem Conclusions-Abschnitt mit mindestens einem „misaligned role code“. [89]

nicht vorhanden wäre. Solche Satzkombinationen sind gängig und sinnvoll. Weniger erwünscht wäre ein viel zu langer Satz, in dem die Definition und die Schlussfolgerung gemeinsam genannt werden. Obwohl ein solcher Satz grammatikalisch komplett korrekt sein kann und von der Bedeutung identisch zu einer Kombination aus zwei Sätzen, würde hier die Lesbarkeit leiden.

„Misaligned codes“ bezeichnet Codierungsstellen bei denen ein Rollen-Code mit einem Inhalts-Code in Verbindung gebracht wird, der dem Codebuch nach nicht zugehörig ist. Insgesamt machen solche Codierungen 5 % aus. Die Studierenden behaupten, dass „misaligned role codes“ sehr selten einzeln in Artikeln auftreten, sondern dazu tendieren würden, gehäuft innerhalb einzelner Artikel aufzutreten. Dies erscheint mir ungenau, wenn man die dazugehörige Grafik (Abbildung 1) betrachtet. Die Grafik umfasst alle 40 Artikel, in denen die Anzahl auftretender „misaligned role codes“ präsentiert werden: In knapp der Hälfte aller Artikel (18/40) treten ein bis zwei „misaligned role codes“ auf, in der anderen Hälfte mindestens drei. Dabei sind von drei bis zehn „misaligned role codes“ die Artikel relativ gleichmäßig verteilt. Es gibt insgesamt nur drei Artikel, die alleine mehr als zehn solcher „misaligned role codes“ haben, wobei zwei davon sogar über 20 aufweisen.

Weitere Ergebnisse dieser Analyse sind, dass background der Code war, der am häufigsten „misaligned“ wurde, gefolgt von method und conclusion-gen. Wie zuvor bereits erwähnt scheint mir die hohe Anzahl an background „misalignments“ von

einer fehlenden Aufmerksamkeit zu den umgebenden Sätzen zu stammen. Diese Annahme wird zusätzlich dadurch gestützt, dass die Mehrheit der Artikel Spezialgebiete behandelt, die für die Studierenden sehr wahrscheinlich unbekannt sind. Die Studierenden erklären „misaligned role codes“ durch unterschiedliche Interpretationen der Code-Definitionen und stützen dies durch eine Grafik, in der die Anzahl „misaligned role codes“ je Coder dargestellt werden. Unter den Codern mit den meisten „misaligned role codes“ findet sich auch mein Name. Das ist plausibel, insbesondere da ich nur zu Beginn des Codierungsprozesses präsent war und nachfolgende Entwicklungen des Codebuchs in meinen codierten Artikeln nicht mehr korrigieren konnte. Allerdings deutet dies letztendlich auf ein unzureichend ausgearbeitetes Codebuch hin, wie auch von den Studierenden selbst vorsichtig angemerkt.

Die Erkenntnisse aus dem „misaligned role codes“ und der Sequenzanalyse stehen etwas gegensätzlich zueinander. Einerseits gibt es anscheinend Artikel in denen Rollen-Codes abwechselnd je Satz codiert wurden, andererseits ist auch bei allein 39 % der „misaligned role codes“ der Code background involviert und bei 20 % der Code method. Das erste deutet auf eine Codierungs-Art hin, bei der einzelnen Sätzen einfach der passende Rollen-Code anhand des zuvor vergebenen Inhalt-Codes zugewiesen wird. Letzteres deutet hingegen darauf hin, dass offensichtlich mehr Wert auf eine einheitliche Sequenz gelegt wurde als darauf, dass der Rollen-Code zum Inhalt-Code passt. Dass Rollen-Codes noch nicht ausgereift sind ist akzeptabel, da diese erst kurz vor dem Start von Prof. Dr. Prechelt aus meiner Erkenntnis, dass Conclusions-Abschnitte ähnlich Abstracts sind, abstrahiert wurden und daher nicht getestet werden konnten. Zusammenfassend lässt sich sagen, dass die Rollen-Codes noch nicht ausgereift sind. Es ist unklar, wie streng Rollen-Codes in Bezug auf Inhalt-Codes codiert werden müssen.

Code-use and non-use: Der Code goodone wurde nie genutzt. Ich erkläre das so, dass es schwerer fällt ein besonders gutes Exemplar zu erkennen als ein besonders schlechtes, vor allem wenn man noch unsicher auf dem Gebiet ist.

Der Code background wurde sehr häufig genutzt (in 52 % der Conclusions-Abschnitte und mehrmals). Dies kann ein Symptom für schlechte Conclusions-Abschnitte sein, aber, wie bereits gesagt, auch unsauberes Anwenden der Kontextregel. Schlussfolgerungen, die als schlecht codiert waren (in 5 % der Conclusions-Abschnitte), waren häufig Schlussfolgerungen mit niedriger Aussagesicherheit, vorenthaltenen Informationen oder wenig neuen Informationen. Insgesamt enthalten nur knapp 80 % aller untersuchten Artikel Schlussfolgerungen. Meiner Annahme nach, dass Conclusions-Abschnitte Schlussfolgerungen enthalten sollten, bedeutet das folgendes: Mindestens 20 % der Artikel erfüllen den Zweck eines solchen Abschnitts nicht. Diese Feststellung bestätigt die Motivation dieser Studie, dass die Qualität der Schlussfolgerungen derzeit unzureichend ist und der Zweck der Schlussfolgerungen in Vergessenheit zu geraten scheint. Die Hypothese von Prof. Dr. Prechelt, dass Autoren sich zunehmend scheuen, Verantwortung für ihre eigenen Schlussfolgerungen zu übernehmen, lässt sich mit den hier verwendeten Methoden nur schwer bestätigen. Wir betrachten lediglich die Artikel eines einzigen Jahres und können daher keinen zeitlichen Vergleich anstellen. Jedoch zeigen die Ergebnisse, dass ein relevanter Teil der Autoren (20 %) das Aufstellen von Schlussfolgerungen in Conclusions-Abschnitten umgeht. Denn laut den Ergebnissen der Studierenden enthalten 99 %

der Conclusions-Abschnitte Ergebnisse, Schlussfolgerungen oder beides, aber nur 80 % enthalten Schlussfolgerungen. Zwar präsentieren die Autoren Ergebnisse, jedoch vermeiden sie das Ziehen von Schlussfolgerungen.

Ich vermute, dass unsichere Autoren nicht immer anhand einer bloßen Häufung unsicherer Formulierungen und somit solcher Codes identifiziert werden können. Autoren priorisieren, wo sie sich am ehesten sicher fühlen, um sowohl sichere als auch unsichere Textstellen zu haben. Sie vermeiden damit möglicherweise eine offensichtlich schlechte Arbeit zu schreiben in der alles unsicher wirkt und damit wahrscheinlich in keiner seriösen Quelle publiziert werden würde. Dadurch entstehen unregelmäßige Conclusions-Abschnitte, in denen Autoren bei zukünftigen Aussagen besonders zuversichtlich sind, um diese fehlende Zuversicht in ihren Schlussfolgerungen auszugleichen. begründen tut dies, dass fposs mit 35 % in den meisten Conclusions-Abschnitten genutzt wurde, gefolgt von fwork (29 %), fhope (21 %), fhope (21 %), fneed (12 %) und fwill (3 %). Auch hier ist kein deutlicher Trend zu unsicheren Schreibweisen zu erkennen. Zwar wird fposs am häufigsten verwendet, aber knapp dahinter fwork. Auch fhope und fshould halten sich die Waage.

Überschriften: 94 % der betrachteten Artikel nutzen die Überschrift „Conclusion“(71 %) oder „Conclusions“(23 %). Die Länge der Artikel mit solchen Überschriften deutet an, dass Artikel unabhängig beider Überschriften ähnlich lang sind. Artikel mit der Überschrift „Conclusions“ tendieren aber zu einer größeren Varianz. Da die Vergabe dieser Codes nicht im Zusammenhang mit anderen Codes betrachtet wurde, lassen sich hier keine Erkenntnisse gewinnen.

Schlussfolgerungen der Studierenden (how to guidelines): Die Schlussfolgerungen der Studierenden beschränken sich streng auf Hinweise zum Schreiben von Conclusions-Abschnitten und treffen kaum Aussagen über andere Ergebnisse, wie dem Stand von Conclusions-Abschnitten. Die Hinweise der Studierenden besagen, dass Schlussfolgerungen mit geringer Konfidenz oder mit nur angedeuteten Informationen mit Vorsicht verwendet werden sollten. Ebenso sollten weniger Übergangssätze verwendet werden, da diese oft als schlechte Sätze gekennzeichnet wurden. Des Weiteren ist es ratsam, keine Ergebnisse zu nennen, ohne darauf folgende Schlussfolgerungen zu präsentieren. Conclusions-Abschnitte sollten niemals ohne Ergebnisse und/oder Schlussfolgerungen erfolgen.

Teile der gegebenen Hinweise stammen ausschließlich vom Code, der schlechte Sätze markiert hat. Dieser Code ist stark subjektiv und beschränkt sich auf „gute und schlechte Beispiele“. Es sei darauf hingewiesen, dass dieser Code nicht zur inhaltlichen Analyse, sondern nur zur Sammlung von Beispielen vorgesehen war. Ich denke, dass goodone und badone trotzdem verwendet wurden, weil die anderen objektiven, den Inhalt beschreibenden Codes, nicht ausreichend sind, um etwas über die Qualität von Schlussfolgerungen herauszufinden, die ausschließlich aus diesen Codes stammen. Die Codes, die dazu dienen, die Qualität zu bewerten, wie comparision oder defensive, waren optionale Codes. Optionale Codes müssen nicht konsequent genutzt werden, und sind daher keine zuverlässige Datenquelle für eine Analyse. Eine mögliche weitere Analyse könnte unsichere und sichere Formulierungen innerhalb eines Artikels untersuchen, worauf das Codebuch jedoch nicht abzielte und auch nicht ausgelegt war. Solche optionalen Codes oder Unterscheidungen gab es konkret

nur für Schlussfolgerungen.

4.2.1 Beurteilung der Ergebnisse der Studierenden

Die Studierenden wiesen darauf hin, dass sich ihr Verständnis der Codierung im Laufe der Zeit geändert habe und dass die Diskrepanzen bei der Codierung daher stammen könnten. Ich schätze, dass dies auch zu Fehlcodierungen geführt hat. Zu beidem hat wahrscheinlich auch beigetragen, dass das Codebuch noch nicht fertiggestellt war. Erschwerend kommt noch dazu, dass Conclusions-Abschnitte tendenziell komplex sind, vor allem wenn zuvor weder Abstract noch die ganze Studie gelesen wurde.

Die Schlussfolgerungen der Studierenden sind eher zurückhaltend. Gemeinsam mit den präsentierten Ergebnissen deutet dies darauf hin, dass das Codebuch in der vorhandenen Form kaum Rückschlüsse auf die Qualität von Conclusions-Abschnitten zulässt. Dies bedeutet auch, dass objektive strukturell inhaltliche Codes, entsprechend der aktuellen Form des Codebuchs, alleine nicht dazu geeignet sind Qualitäts-Faktoren zu bestimmen.

Die Codierungen der Studierenden scheinen im Wesentlichen sinnvoll zu sein. Bei dem Vergleich meiner codierten Artikel mit denen der Studierenden (40 verglichene Artikel), gab es ungefähr drei Unstimmigkeiten pro Artikel. Da nicht alle der Unstimmigkeiten immer Fehlcodierungen sind, sondern teilweise auch akzeptable verschiedene Interpretationen, schätze ich, dass die Ergebnisse der Studierenden annehmbar sind und als Tendenzen verstanden werden sollten. Dennoch habe ich einige problematische Stellen gefunden, als ich meine eigenen Codierungen vom ersten Artikel an mit der letzten Version des Codebuchs (Ende Juli) neu codiert habe:

- Das Codebuch war nach der offenen Codierung noch nicht ausgereift. Einige Codes wie fneed und conclusions wurden häufig bei gleichen Sätzen genutzt. Die Definitionen der Codes waren während des Projekts nicht differenzierend genug. Dadurch kam es zu ungenauen Codierungen.

- likeabstract wurde selten verwendet und nicht konsequent, da es ein optionaler Code war. Möglicherweise wurde der Code auch inkonsequent angewandt, da er zu ungenau definiert war. Er lässt sich daher nicht gut zur Auswertung nutzen.

- In manchen Conclusions-Abschnitten wurden Sätze in Isolation codiert anstatt im Kontext. Dies ist bei folgendem Beispiel ersichtlich. Der folgende Satz wurde von Studierenden als SS und relatedwork codiert. „Few SE studies use random sampling“ [4]. In Isolation wären das passende Codes, im Kontext des ganzen Conclusion-Abschnitts passt conclusion-gen besser. Denn Forschungsgegenstand des Artikels war eine begrenzte Menge an SWE-Artikel von denen aus nun auf die gesamte Menge an SWE-Artikeln geschlossen wird. Dies geht hervor durch das zuvor genannte objective der Studie: „An analysis of the state of sampling methods in a stratified random sample of recent SE research in leading venues“ [4]. Die Studierenden haben nicht berücksichtigt, was vor und was nach der Studie bekannt war. Solche Fehler traten häufig bei systematischen Literaturrecherchen auf.

Background wurde außerdem manchmal an unpassenden Stellen wie dem OC codiert, da er dort als Definition für eine Schlussfolgerung benötigt wurde. Durch häufige Vergabe fehlerhafter Codes sind vermehrt „misaligned role codes“ entstanden. Die Menge an „misaligned role codes“ scheint daher wenig aussagekräftig zu sein und zeigt höchstens Tendenzen an.

- Der Code teaser wird nicht konsequent vergeben, wahrscheinlich weil dies ein optionaler Code ist. Außerdem erfordert der Code eine andere Denkweise als das simple Einordnen von Inhalt. Eine Auswertung mit diesem Code scheint daher nur begrenzt möglich.
- Auch abbrev wird inkonsequent vergeben und fehlt daher an einigen Stellen.

Mit den Ergebnissen der Studierenden ist deutlich geworden, dass es anscheinend nicht sinnvoll ist, die Qualität von Conclusions-Abschnitten mit dem bisherigen Struktur fokussierten Codebuch zu beurteilen. Auch ist deutlich geworden, dass das Codebuch noch einige unfertige Stellen hat.

4.2.2 Reflexion zum Studierendenprojekt

Insgesamt hat das Studierendenprojekt gut funktioniert. Die Studierenden haben motiviert gearbeitet und waren bereit, sich mit offenem Codieren, dem Codebuch und der Qualität von Conclusions-Abschnitten auseinander zu setzen. Das Ergebnis des Projekts ist hilfreich, da es Gewissheit gibt, dass das Codebuch mit fast ausschließlich strukturell inhaltlichen Codes nicht allein zur Beurteilung der Qualität ausreichen kann.

Durch die Vorlage eines halbfertigen Codebuchs konnte Inhaltsanalyse besser beispielhaft erlernt werden. Es konnte viel Zeit beim offenen Codieren gespart werden. Trotzdem gab es ein nutzbares Codebuch und der Grundgedanken der offenen Codierens konnte gezeigt werden.

Die Nutzung, der von Prof. Dr. Prechelt für die für qabstract-Studie eigens erstellten Codierungsumgebung brachte Vor- und Nachteile mit sich. Ein großer Vorteil war, dass die Studierenden kaum organisatorischen Aufwand hatten. Artikel konnten einfach codiert und mit der Version der anderen codierenden Studierenden einfach abgeglichen werden.

Als Nachteile zeigten sich: Erstens, dass das Vorbereiten des Codebuchs war eine persönliche Herausforderung, da es an die neue von der qabstracts-Studie übernommene Codierungsumgebung angepasst werden musste. Diese unterschied sich von MAXQDA. In MAXQDA kann mit beliebiger und wechselnder Granularität codiert werden und das Codieren wird automatisch unterstützt durch z.B. Vervollständigung der Codebenennung. Im Vergleich zur Codierungsumgebung des Studierndenprojektes ist das Codieren in MAXQDA wesentlich freier. In der Codierungsumgebung des Projektes sollte jeder Satz genau einem Code entsprechen. Außerdem ist die Lesbarkeit des Textes in der Umgebung schlechter als in MAXQDA, da Absätze nicht angezeigt werden und Codes direkt im Text festgehalten werden. Deshalb habe ich mich bemüht, meine MAXQDA-Codes anzupassen. Dafür wurden alle Codes auf Satzgranularität angepasst, Codebezeichnungen gekürzt und ich begann zwischen den immer zu vergebenden Codes und optionale Codes bzw. Suffixen zu unterscheiden.

Beim Codeabgleichen mit den Studierenden sind mir mehrere Dinge aufgefallen. Das direkte Codieren im vorgegebenen Format war nicht so, wie offenes Codieren sein sollte, da dadurch gleich von Beginn an eine Granularität vorgegeben wurde. Allerdings gab es bereits ein angefangenes Codebuch mit einer festen Granularität und dies sollte verstanden werden. Das Codieren war somit passend zum Stand des Codebuchs. Es gab Codeabgleiche bei denen ich mir sicher bin, dass die Studierenden viel mitgenommen haben und mit einem besseren Codeverständnis herausgingen. Bei solchen Abgleichen konnte ich auch lernen. Andere Codeabgleiche waren eher mechanisch und es gab kaum Interaktion und Lernmöglichkeiten. Codeabgleiche waren immer besser, wenn Studierende das Skript zum Auffinden von Codierungskonflikten nicht nutzten. Diese Art des Abgleichs führte dazu, dass der Kontext des Textes weniger beachtet wurde und generell weniger Diskussionen zu Stande kamen, bei denen etwas über die Codes hätte gelernt werden können. Für eine Phase, in der ein Codebuch erlernt werden sollte und möglichst viel Austausch stattfinden sollte, ist es ratsam, per Hand abzugleichen. Für eine quantitative Codierungsphase, in der das Codebuch weitestgehend beherrscht wird und möglichst viele Artikel codiert und verglichen werden sollten, ist das Skript gut geeignet.

Durch die Art des Codierungsprozesses des Studierendenprojekts und der Darstellung des Textes gab es kaum Anreize, den Conclusions-Abschnitt in seiner Gesamtwirkung zu betrachten. In meiner weiteren Forschung hat mir das allerdings geholfen. Den ganzen Artikel als Ganzes zu betrachten, ist wichtig für die Beurteilung der Qualität von Conclusions-Abschnitten, da sie auf diese Weise auf Lesende wirken. Ein insgesamt gut erscheinender Conclusions-Abschnitt kann auch schlechte Qualitäts-Faktoren beinhalten und umgekehrt, da Qualitäts-Faktoren unterschiedlich gewichtet sein können. Würde der Kurs erneut in dieser Form stattfinden, würde ich vorschlagen, codierte Conclusions-Abschnitte auch in ihrer Gesamtheit mit einem Code zu versehen, der eine allgemeine Abschätzung der Qualität dieser beinhaltet.

Aufgrund der Art, wie Conclusions-Abschnitte extrahiert wurden, sind Text-Strukturen verloren gegangen. D.h. Absätze als Trennung zwischen Satzblöcken und Aufzählungen in Form von Listen sind verschwunden. Dadurch geht teilweise Textkontext verloren in Bezug auf Zusammenhänge zwischen den Sätzen. Besonders unschön zu lesen waren Conclusions-Abschnitte, die mit Listen strukturiert waren. Diese Art der Strukturierung wurde in der Codierungsumgebung des Studierndenprojekts nicht dargestellt und ließ den Text eher schwer lesbar und fragmentiert erscheinen. Damit hatten Listen einen gegenteiligen Effekt auf die Lesbarkeit des Abschnitts.

Würde das ganze Projekt nochmal durchgeführt werden, sollte mehr Zeit für das offene Codieren eingeplant werden. So wie es hier durchgeführt wurde, mussten die Studierenden in derselben Woche das neue Codebuch erlernen und Codebuchverbesserungen entwickeln. Und bereits in der Folgewoche sollte das Codebuch für die folgende quantitative Analyse fertiggestellt sein. Ein Tutorium fiel aus, weshalb es nicht möglich war, die Codeverbesserungen in der Gruppe abschließend zu diskutieren und abzustimmen. Die Verwendung von Discord und Abstimmungen per Icons war zwar akzeptabel, aber meiner Meinung nach hätte ein persönlicher Austausch die Lernprozesse der Studierenden verbessert. So wären die Änderungen gleich allen Studierenden bekannt gewesen und die Diskussion hätte möglicherweise noch bessere Codebuchverbesserungen erbracht. Für das letzte übertragen von Codebuch-

verbesserungen sollte ein Tutorium genutzt werden, so dass alle Studierenden sich aller Codebuchveränderungen bewusst wären und diese auch final nochmal diskutiert werden könnten.

Das Codebuch war nach dem Abschluss der Phase des offenen Codierens nicht abgeschlossen. Um dies in einem Folgeprojekt zu vermeiden, könnte den Studierenden mehr Zeit für das offene Codieren gegeben werden, um mehr Codebuchverbesserungen zu erzeugen oder das initial präsentierte Codebuch könnte ein fertigeres Codebuch sein.

Rückblickend scheint es so, als ob der Fokus zu sehr auf strukturellen Inhalt-Codes lag. Da ich für meine eigene Forschung nach dem Studierendenprojekt festgestellt habe, dass die Strukturcodes zwar nicht besonders hilfreich sind, um die Qualität des Conclusion-Abschnitts festzustellen, aber sehr hilfreich sind, um sich innerhalb des Artikels zu orientieren, schlage ich vor, die offene Codierungsphase ohne Codebuch zu beginnen, um den Studierenden möglichst viel Freiraum zu geben. Dann sollte das Codebuch mit den inhaltlichen Codes eingeführt werden und mit den von den Studierenden entwickelten Codes abgeglichen und ergänzt werden. Da dies keine Tätigkeit ist, in der offenes Codieren geübt wird, könnte der Tutor dies übernehmen und den Studierenden dann präsentieren. Mit Hilfe des strukturell-inhaltlichen Codebuchs sollten die Studierenden in der Lage sein, innerhalb des Artikels schnell interessante Stellen zu erkennen, die dann genauer betrachtet werden können. Dabei wäre eine geführte Codierung mit einem Beispieltext, einer Musterlösung und verschiedenen Codes oder zusätzlichen Code-Dimensionen je nach Betrachtungsperspektive hilfreich. Mit den ausdefinierten inhaltlichen Struktur Codes würde den Studierenden eine Basis gegeben werden mit der sie sich in Artikeln zurechtfinden könnten. Damit müssten sie diese inhaltlichen Struktur Codes nicht mehr fokussieren, sondern nur noch anwenden und könnten sich damit auf andere vielversprechendere Perspektiven, wie z.B. die rhetorische Perspektive fokussieren können.

Ein weiterer Faktor, der sicherlich nicht förderlich war, aber auch nicht zu beheben sein wird, ist die Motivation der Studierenden. Die Studierenden arbeiteten unterschiedlich engagiert mit. Der Unterschied von dem, was im Tutorium diskutiert wurde, und dem, was außerhalb des Tutoriums gemacht und entwickelt wurde, war sichtbar. Während des Tutoriums beteiligten sich alle Studierenden an den Diskussionen und erarbeiteten dabei verschiedene Ansätze für dasselbe Thema. Da der Kurs nur einmal pro Woche ein Tutorium hat, bei dem alle anwesend sind und sonst asynchron stattfindet, sind Diskussionen außerhalb des Tutoriums schwer umzusetzen. Dadurch unterscheiden sich die endgültigen Lösungsvorschläge kaum von den ursprünglichen Vorschlägen. Eine mögliche Verbesserung wäre, dass jeder Einzelne sich aktiv an der Online-Diskussion beteiligen muss und nicht jede Gruppe. Es könnte etwas zu viel verlangt sein, jeder Gruppe in der letzten Woche des offenen Codierens einen zehnminütigen Vortrag aufzuerlegen, in dem sie Verbesserungsvorschläge vorstellen und begründen sollten. Dies könnte jedoch dazu führen, dass die Studierenden sich nur mit der Erstellung der Vorschläge beschäftigen, ohne sie in der Gruppe ausführlich zu diskutieren. Außerdem ist der zeitliche Aufwand so zu halten, dass er der Modulbeschreibung weiterhin entspricht.

4.3 Erkenntnisse aus dem Studierndenprojekt

Folgende Erkenntnisse konnte ich aus dem Projekt der Studierenden mitnehmen:

- Häuft sich innerhalb eines Artikels Ungewöhnliches zu oft an, wird man als Codierer dem gegenüber blind. Teilweise durch die Art der Forschung, welche ähnlich einer Literaturrecherche war, und teilweise weil die Autoren damit ihre Argumente belegten. Dies führte dazu, dass ich (und der Student) beim Codieren irgendwann so daran gewöhnt waren, Sätze zu sehen die ausschließlich verwandte Arbeiten besprachen, dass wir es nicht mehr als etwas Ungewöhnliches wahrnahmen, das es zu codieren galt.
- Gut geschriebene Artikel scheinen weniger Schwierigkeiten bei der Codierung zu verursachen und sind einfacher zu lesen. Außerdem führen sie zu weniger Diskrepanzen bei der Codierung, zumindest in der begrenzten Menge, die ich mir angesehen habe.
- Schlussfolgerungen sind eher Ratschläge, die man direkt nutzt, weniger Empfehlungen für die Zukunft.
- Codes, die den strukturellen Inhalt beschreiben sind nicht geeignet um die Qualität eines Conclusions-Abschnittes zu beurteilen.

4.3.1 Erkenntnisse zum Codebuch

Das Codebuch enthält fast ausschließlich nur objektive Codes, die den strukturellen Inhalt eines Satzes innerhalb eines Conclusions-Abschnitts beschreiben. Ein Grund dafür ist, dass dass die Codes für eine quantitative Studie zur Qualität von Schlussfolgerungen möglichst objektiv sein sollten. Daher konzentrierte ich mich auf Codes, die von vornherein objektiv waren, und vernachlässigte subjektive Codes so umzukonzipieren, dass sie von objektiven Codes beschreibbar waren. Ein weiterer bereits erwähnter Grund war, dass ich qualitatives Analysieren gerade erlernte und daher noch nicht in der Lage war, gezielt zwischen Gedankenebenen beim Codieren zu unterscheiden. Ein dritter Grund war, dass mich die qabstracts-Studie und das Codebuch dazu beeinflusst haben. Denn diese Studie diente als Vorlage für mein Codebuch. Da ich zuvor keine Erfahrung mit qualitativer Forschung hatte, habe ich mich wahrscheinlich zu sehr auf die qabstracts-Studie als Ziel fokussiert. Dass es fast ausschließlich nur objektive Codes geben sollte, beruhte neben der verringerten Komplexität auch auf der Hoffnung, dass es mit diesen möglich sein könnte, die Qualität von Conclusions-Abschnitten zu beurteilen. Die Beurteilung der Qualität, beruhend auf fast ausschließlich objektiven Kriterien, wäre um einiges vorteilhafter, als wenn mehr subjektive Kriterien beteiligt wären. Dass dies nicht möglich ist, wurde erst durch die Ergebnisse der quantitativen Studie der Studierenden offensichtlich. Daher der Wechsel zum Vergleichen subjektiver Extreme.

Bei der Nutzung des Codebuchs durch die Studierenden sind einige unklare Stellen im Codebuch aufgezeigt worden, beschrieben in den ersten fünf folgenden Listeneinträgen. Auch in der nachfolgenden Zeit sind Probleme im Codebuch klarer geworden. Zwecks Lesbarkeit sind hier alle aufgefallenen Stellen, die verbesserungswürdig sind, genannt:

- Die Anwendung der Rollen-Codes müsste festgelegt werden. Unklar ist, ob Rollen-Codes eher folgend auf vergebene Inhalt-Codes vergeben werden sollten oder umgekehrt. Auch sollte in diesem Zusammenhang geklärt werden, wann „misaligned role codes“ akzeptabel sind und wann nicht.

- Noch immer sind die Grenzen zwischen SS und OC nicht eindeutig definiert, so auch zwischen OC und OL.

- Wenn optionale Codes optional sind, aber die verpflichtenden Inhalt-Codes kaum etwas über die Qualität aussagen, muss überlegt werden, ob die optionalen Codes auch verpflichtend werden. Wahrscheinlich müssen noch weitere Codes gefunden werden.

- Manche Sätze lassen sich ohne Vorwissen aus dem Artikel schwer codieren. Bereits codierte Texte erscheinen bei erneuter Betrachtung teils unpassend codiert. Es ergeben sich Codierungsoptionen, die sich aus Beziehungen zwischen Sätzen und dem wahrscheinlich fehlendem Wissen ergeben, was gezwungener Maßen durch Spekulation ersetzt wird. Ein Code der solche Unsicherheiten markiert wäre hilfreich, um darauf aufmerksam zu machen, dass solche Stellen vorsichtig zu analysieren sind. Innerhalb MAXQDAs habe ich den Code hardtocode dafür eingeführt.

- Das Codieren mit Kontext erzeugt Konflikte. Folgende Codierungsweise wurde, den vergebenen Codes nach, nicht nur von mir genutzt: Wenn ich eine Schlussfolgerung codiere und daraufhin ein Resultat folgt, das die Schlussfolgerung erklärt, lasse ich das Resultat als solches codiert. Wenn dasselbe mit einem Satz passiert, der ohne Kontext als Hintergrund codiert werden würde, dann wird dieser Satz als Schlussfolgerung codiert. Die Annahme dahinter ist, dass Hintergrundinformationen nur in der Einleitung erwünscht sind und sonst nicht, während Ergebnisse logischerweise in Verbindung mit Schlussfolgerungen erwartet werden. Diese Lösung erscheint mir jedoch nicht ideal. Durch diese Vorgehensweise lässt sich nicht zwischen results, die direkt Details zu Schlussfolgerungen liefern, results, die die Studienergebnisse allgemein beschreiben, und results, die selber fast Schlussfolgerungen sind, unterscheiden. Manche Autoren vermeiden es zu Generalisieren und führen stattdessen fast generalisierte Resultate an, die als Schlussfolgerungen gemeint sind. Zwischen diesen Arten von Resultaten sollte unterschieden werden, da sie unterschiedliche Arten an Informationen tragen und zu unterschiedlichen Rollen beitragen und damit die Qualität von Conclusions-Abschnitten verschieden beeinflussen. Resultate, die fast Schlussfolgerungen sind oder Detailinformationen zu Schlussfolgerungen gebenm, sind eher Teil der Studienerkenntnisse (OC) und nicht der Studienbeschreibung (SS).

Sollte Interesse daran bestehen, das Codebuch weiter zu nutzen, empfehle ich das Codebuch als eine Art Orientierungshilfe innerhalb von Conclusions-Abschnitten zu verstehen und qualitätsbeschreibende Merkmale mit einer eigens erneut angewandten qualitativen Analyse zu finden.

4.3.2 Aufbau von Conclusions-Abschnitten

Conclusions-Abschnitte sind nach den Studierenden in den meisten Fällen ca. 200–300 Wörter lang und folgen zu ca. 80 % einer Abstracts ähnelnden Struktur, die ungefähr dem chronologischen Ablauf einer Studie folgt. Ein Abstract beginnt immer mit dem Hintergrund und der Motivation der Autoren zur Forschung, erläutert dann deren konkreteres Ziel, die genutzte Methodik um dies zu erreichen, nennt im Anschluß die wichtigsten Ergebnisse und zieht dann daraus Schlussfolgerungen und einen Ausblick. Davon gibt es zwei Varianten, der komplett lineare Conclusions-Abschnitt und der themengruppierte Conclusions-Abschnitt. Jeder Conclusions-Abschnitt folgt grob der Struktur eines Abstracts. Ein *linearer Conclusions-Abschnitt* tut dies strikter: Ein Conclusions-Abschnitt beginnt mit OR oder direkt mit der eigentlich nachfolgenden Studienberschreibung (SS). Es folgen SS, OC und OL. Beide Ausprägungen machten jeweils 40 % der analysierten Artikel des Studierendenprojekts aus. Ein linearer Conclusions-Abschnitt zählt innerhalb des OC erst die Resultate auf, dann alle daraus folgenden Schlussfolgerungen und zuletzt gesammelt den Ausblick zu allen Erkenntnissen. Ein Beispiel dafür ist Abbildung 4 auf Seite 69.

Der themengruppierte Abschnitt unterscheidet sich darin, dass dieser mehr nach den Themen der einzelnen Schlussfolgerungen sortiert ist. Das bedeutet, dass Informationen bezüglich einzelner Schlussfolgerungen um diese herum gruppiert sind. Für jede Erkenntnis werden einzeln die relevanten Resultate und ggf. Hintergrundinformationen erläutert, so wie die folgenden Implikationen gegeben. Dies wird für jedes Thema der Studie wiederholt. Theoretisch könnte dies dazu führen, dass einzeln auch klein gefasste Studienziele oder Methoden direkt neben Schlussfolgerungen auftauchen könnten. Solch einen Conclusions-Abschnitt habe ich allerdings nicht gesehen. Informationen, die alle Schlussfolgerungen betreffen bzw. allgemeiner sind, wie das generell Studienziel, verbleiben gesammelt an derselben Stelle, wo sie auch in einem linearen Conclusions-Abschnitt gefunden werden würden. Ein Beispiel hierfür ist Abbildung 2 auf Seite 67.

Unterscheidungen der Ziele von Conclusions-Abschnitten: Es scheint, als ob Conclusions-Abschnitte je nach Autoren andere Ziele verfolgen und dementsprechend den Lesenden verschiedene Informationen vermitteln. Folgende Beschreibungen sind eigene Vermutungen, keine belegten Erkenntnisse. Sich wiederholende Ziele scheinen folgende zu sein. Es können mehrere Ziele innerhalb eines Abschnitts abgedeckt werden:

- **Eine Zusammenfassung wie ein Klappentext.** Es wird beschrieben, was herausgefunden wurde, ohne die konkreten Informationen dabei zu nennen. Damit bleiben Lesende eines solchen Abschnitts unwissend bezüglich der Erkenntnisse der Studie. Möglicherweise sollen Lesende so motiviert werden, den Hauptteil der Studie zu lesen. Denn meist werden die Erkenntnisse in einer Art und Weise beschrieben, die deren Relevanz und Bedeutung positiv hervorhebt.

- **Weitergeben der wichtigsten Erkenntnisse.** Diese Art von Conclusions-Abschnitten erfüllen den angedachten Zweck, indem sie die Schlussfolgerungen

präsentieren. Meist verlieren solche Abschnitte nicht besonders viel Zeit damit, ausführliche Hintergrundinformationen zu geben.

- **Das Abstract wiederholen.** Diese Abschnitte gleichen Abstracts in der Struktur und darin, dass sie eher kurz gehalten sind und jeden Teil der Studie zusammenfassen. Solche Conclusions-Abschnitte enthalten kaum neue Informationen im Vergleich zum Abstract. In manchen Fällen sind die Wortlaute nahezu identisch.

Die Überschriften der Conclusions-Abschnitte selber lassen kaum Rückschlüsse auf den Inhalt des Abschnitts zu. Weder „Conclusions“ noch „Conclusion“ deuten auf Schlussfolgerungen bzw. keine Schlussfolgerungen im Abschnitt hin. Einzig ungewöhnliche Überschriften mit Wörtern, wie „Implications“ oder „Learned Lessons“, beschreiben den Inhalt des folgenden Conclusions-Abschnitt verlässlich.

Ähnlich verhält es sich mit den Gliederungen von Artikeln, die, wenn vorhanden, am Ende des Einleitungs-Abschnitts zu finden sind. Von den 48 Artikeln, die ich analysiert habe, haben 34 eine Gliederung am Ende des Einleitungs-Abschnitts. Davon versprechen 12 Artikel im letzten Abschnitt „conclusions“ und „learnings“, aber nur die Hälfte der Abschnitte enthält konkrete Schlussfolgerungen. 15 Artikel beschreiben den Inhalt ihres letzten Abschnitts als „abschluß der Studie“ („conclude the study“). Diese Beschreibung ist so weit gefasst, dass sich von guten Conclusions-Abschnitten mit Fokus auf Schlussfolgerungen, über kurze Zusammenfassungen und lange Texte ohne Conclusions alle Arten von Conclusion-Abschnitten wiederfinden lassen. Sieben der Artikel beschreiben den Inhalt des letzten Abschnitts genauer und verwenden ungewöhnlichere Beschreibungen, wie z.B. „[The last section] concludes by discussing our findings, the limitations of the study, and an outlook to future work“ [6] oder „[The last section] discusses the learned lessons and our future work“ [24]. Sechs der Conclusions-Abschnitt dieser Artikel enthalten konkrete Schlussfolgerungen und gehören zu den guten bzw. eher guten Conclusions-Abschnitten, nur einer der Abschnitte ist eher schlecht. Beschreibungen des letzten Artikels, die ungewöhnlich erscheinen und eher seltener zu sehen sind, beschreiben den Inhalt des Abschnitts mit hoher Zuverlässigkeit genau. Sofern die Beschreibung auf Schlussfolgerungen hinweist, sind diese auch fast immer gegeben.

Fazit zum Aufbau von Conclusions-Abschnitten: Dass die meisten Conclusions-Abschnitte strukturell Abstracts ähneln, scheint mir akzeptabel. Der Aufbau von Abstracts ist bekannt und Lesende finden sich daher schnell zurecht darin. Allerdings sind nicht alle in Abstracts enthaltenen Informationen relevant für das Teilen der Schlussfolgerungen. Dies betrifft vor allem die Motivation, so wie die Hintergrundinformationen der Studie. Beides sind Informationen, die im Abstract bereits zur Genüge besprochen wurden und weder die Schlussfolgerung selber beschreiben, noch deren Validität oder Relevanz verbessern. Motivation und Hintergrundinformationen beschreiben die Relevanz der Forschung an sich, nicht aber die der Ergebnisse. Damit sind Sätze mit solchem Inhalt in einem Conclusions-Abschnitt überflüssig. Das bedeutet, dass ca. 40 % aller Conclusions-Abschnitte unnötige Informationen enthalten.

Die Studienbeschreibung (SS) mit Methoden und Resultaten dagegen ist eine Möglichkeit, die Validität und Relevanz der Erkenntnisse der Studie zu beschreiben

und die Schlussfolgerungen glaubwürdiger erscheinen zu lassen. Da manche Lesende Conclusions-Abschnitte direkt nach dem Abstract lesen, können die, wenn nicht identisch zum Abstract, erwünschte Informationen enthalten. In allen Fällen zuträglich ist es, das Ziel der Studie zu wiederholen (auch Teil von SS) und dieses damit Lesenden präsenter zu machen und die Möglichkeit zu haben, auf dieses Bezug zu nehmen. Der wichtigste Teil des Conclusions-Abschnitts sind die Schlussfolgerungen selber (enthalten in OC). Damit ist ein OC-Teil in jedem Conclusions-Abschnitt zu erwarten. OL-Teile sind die Folgen der Erkenntnisse und damit ebenfalls verpflichtend für jeden Conclusions-Abschnitt.

Die Studierenden haben herausgefunden, dass nur 20 % von Conclusions-Abschnitten Schlussfolgerungen enthalten. Meine Überprüfung der codierten Segmente der Studierenden ergab Codierungskonflikte bei ca. 10 % der als conclusion codierten Sätze. Allerdings tendieren die Studierenden eher dazu, häufiger den Code conclusion zu vergeben, weshalb die festgestellten 20 % von Abschnitten ohne Schlussfolgerung mehr sein könnten. Mindestens jeder fünfte Artikel enthält damit einen schlechten Conclusions-Abschnitt. Wahrscheinlich ist die Rate an schlechten Conclusions-Abschnitten sogar größer, da ich während meiner Forschungsarbeit auch schlechte Conclusions-Abschnitte gesehen habe, die Schlussfolgerungen enthielten. Conclusions-Abschnitte gibt, die mit Motivation und Hintergrundinformationen beginnen (OR). Solche die dies ausführlich tun, nehme ich häufiger als schlechter wahr, als solche, die dies nicht tun. Motivation und Hintergrundinformationen sollten bereits im Abstract ausreichend erläutert worden sein, welches i.d.R. von Lesenden des Conclusions-Abschnitts gelesen wurde. Diese Informationen sind für den Ausgang der Studie wenig relevant und haben es daher nicht nötig, kurz vor den Schlussfolgerungen wiederholt werden zu müssen. Anders sieht dies mit der Studienbeschreibung aus, die zwar auch im Abstract enthalten sein sollte, aber bei der es sinnvoll ist, das Ziel der Studie zu wiederholen und den Lesenden ins Gedächtnis zu rufen, um dieses dann mit den Schlussfolgerungen im OC begründen zu können. Schlussfolgerungen werden dabei meist nicht alleine genannt, sondern häufig in Verbindung mit deren zugrunde liegenden Resultaten, welche auch Teil des OC sind. Beendet werden Conclusions-Abschnitte beider genannter Strukturen mit dem Ausblick (OL).

5 Suche nach Qualitätsfaktoren

Nach dem Studierendenprojekt war klar, dass objektive, inhaltliche Codierungen nicht besonders hilfreich sind, um die Qualität von Conclusions-Abschnitten zu bestimmen. Daher bin ich dazu übergegangen, Extreme miteinander zu vergleichen. Diese Extreme beruhen auf meiner subjektiven Einschätzung darüber welche Conclusions-Abschnitte gut und welche schlecht sind. Deswegen habe ich im Anschluss eine kleine informelle Studie durchgeführt, die in Erfahrung bringen sollte ob meine auf subjektiven Einschätzung beruhenden Qualitätsfaktoren gültig sind.

Das folgende Kapitel ist so aufgebaut, dass es zuvor genannte Forschungsabschnitte beschreibt und dabei jeweils einen Bericht über Geschehnisse und Resultate liefert und eine Zusammenfassung inklusive Selbtsreflexion. Nach den Forschungsergebnissen folgen die detaillierten Erkenntnisse bezüglich der Qualität von Conclusions-

Abschnitten.

5.1 Vergleichen von Extremen

Auf Empfehlung von Prof. Dr. Prechelt habe ich Extreme miteinander verglichen. Ich begann mit Sätzen von Schlussfolgerungen, die mir persönlich entweder besonders gut oder besonders schlecht erschienen. Um die Anzahl an Sätzen zu erhöhen, codierte ich Conclusions-Abschnitte in denen Studierende besonders gute bzw. schlechte Schlussfolgerungen codiert hatten: conclusion-gen:low, goodone, badone.

Dabei habe ich folgende Eigenschaften gefunden, die die Qualität von Schlussfolgerungen senken:

- Schlussfolgerungen, die nur beschrieben, aber nicht direkt präsentiert werden senken die Verständlichkeit und damit die Qualität des Schlussfolgerungssatzes.
- Schlussfolgerungen, die so formuliert sind, dass die Aussage aufgeweicht wird oder unkonkreter wird sehe ich als schlecht an. Verschiedene Formulierungen sind unterschiedlich schlecht. Eine detaillierte Erläuterung ist in Abschnitt 5.3.2 zu finden.
- Das Auslassen von Generalisierungen führt häufig zu einer Aufzählung an Resultaten aus denen der Leser eigene Schlussfolgerungen ziehen soll: Dies ist eher ungünstig, da Lesende normalerweise weniger Informationen zur Verfügung haben als Autoren und Autoren sich vor der Verantwortung drücken, Schlüsse aus ihrer eigenen Arbeit zu ziehen. Ausnahme sind hier Forschungsergebnisse, die sich nicht generalisieren lassen. Dies soltle dann allerdings enstprechend begründet werden im Conclusions-Abschnitt.
- Das Negieren von Generalisierungen: „We do not claim the generalization of our observed results. Instead, we simply observe that there exist software projects in which using SNA metrics by themselves or along with code metrics could yield better performing SDP models across several SDP contexts and scenarios“ [34]. Obwohl mir der Zweck solcher direkt negierten Generalisierungen mir unklar ist und dies offensichtlich auch nicht dem Zweck von Conclusions-Abschnitten und dem Bilden von Schlussfolgerungen entspricht, wird das Wissen bei negierten Schlussfolgerungen doch relativ offensichtlich weitergegeben.
- Zu lange Sätze erschweren das Verständnis des Satzes. Das folgende Beispiel zeigt einen solchen langen Satz, der bei ungeübten Lesenden dazu führt, dass der Satz mehrmals gelesen werden muss: „More specifically, the methodology was able to identify factors that were influential in decision-makers' choices between software technologies and demonstrate the need to account, not only for the properties of the choice, but also of the chooser and of the importance of the supply chain dependencies and information flows“ [56].

Bei der weiteren Untersuchung der Merkmale kam ich wieder auf die Verbesserung meiner bestehenden strukturellen Inhalts-Codes zurück. Um mich von den Struktur-Codes zu lösen, versuchte ich, für jeden meiner strukturellen Codes herauszufinden, was dieser Code über die Qualität der Schlussfolgerungen aussagen könnte.

Diese Aufgabe hatte ich im Verlauf meiner Forschung schon zu verschiedenen Gelegenheiten begonnen, war aber nie zu einem befriedigenden Ergebnis gekommen. Dies schrieb ich zu Beginn meinem noch nicht ausreichenden Wissen zu Schlussfolgerungen und Conclusions-Abschnitten zu. Mittlerweile war ich mir aber sicher, genug theoretische Sensitivität angeeignet zu haben. Trotzdem erhielt ich keine Ergebnisse, mit denen ich das Gefühl bekam mit Hilfe struktureller Codes Schlüsse zur Qualität ziehen zu können. Im Kontext der Ergebnisse der Studierenden wurde mir klar, dass ich nie zu befriedigenden Ergebnissen gekommen war, weil die Codes dies einfach nicht hergaben. Die Qualität von Conclusions-Abschnitten lässt sich nicht durch strukturell inhaltliche Codes beurteilen.

Im folgenden Arbeitsschritt verglich ich komplette Conclusions-Abschnitte miteinander. Die Festlegung, welche Abschnitte besonders gut oder besonders schlecht waren, beruhte auf meinem subjektiven Lesempfinden des jeweiligen Artikels. Dafür habe ich alle codierten Conclusions-Abschnitte (48) in vier Kategorien eingeteilt: schlecht, eher schlecht, eher gut, gut. Insgesamt ergab dies sieben gute und acht schlechte Conclusions-Abschnitte. Vorgegangen bin ich so, dass ich gute und schlechte Abschnitte uncodiert ausgedruckt habe und dann gelesen habe. Dabei habe ich grob OR, SS und OL markiert und habe mich nur näher mit OC befasst. Anschließend las ich alle Artikel erneut und notierte dabei alles, was einen guten und einen schlechten Eindruck verursachte. Dabei habe ich darauf geachtet, welche Sätze mir schwer fielen zu lesen oder zu codieren. Ich habe mich jeweils gefragt, warum das so war. Zusätzlich habe ich immer versucht, die Ergebnisse der Conclusions-Abschnitte zusammenzufassen und zu überlegen ob ich als Leserin das Gefühl hatte, dass mir neues Wissen präsentiert wurde.

Aus dieser Analyse habe ich gelernt, dass meine subjektive Bewertung von guten und schlechten Conclusions-Abschnitten stark vom Vorhandensein von Schlussfolgerungen in den Abschnitten abhängt und dass die Präsentation dieser weniger Einfluss hat, solange Schlussfolgerungen als solche erkennbar bleiben. Das bedeutet, dass Schlussfolgerungen gemacht werden müssen und nicht nur durch Ergebnisse angedeutet werden sollten. Vorhandene Schlussfolgerungen müssen sowohl ausreichend konkret und lesbar sein als auch ausreichend Platz im Abschnitt einnehmen, damit es ein guter Conclusions-Abschnitt ist. Alle Informationen müssen vollständig ausgeführt werden. Es ist auch wichtig, dass die Schlussfolgerungen eindeutig erkennbar sind, zum Beispiel durch persönliche Empfehlungen des Autors. Die Glaubwürdigkeit der Schlussfolgerungen scheint eine weitere offensichtliche Dimension zu sein, aber diese hängt von der Forschung selbst ab und von der Art und Weise, wie sie präsentiert wird, d.h. praktisch von der Argumentation der Forschung, die ohnehin das Hauptthema jedes Artikels ist. Eine ausführliche Liste aller Einflüsse findet sich in Kapitel 5.3.

Nachdem ich die Extreme verglichen hatte, habe ich die eher guten bzw. schlechten Conclusions-Abschnitte in die Analyse miteinbezogen, um das Zusammenspiel der gefundenen Qualitätsfaktoren besser zu verstehen. Dies ergab keine besonders detaillierten Zusammenhänge, sondern bestätigte hauptsächlich die zuvor bereits genannte Theorie, dass die Qualität von Conclusions-Abschnitten hauptsächlich von dem Informationsgehalt an Schlussfolgerungen abhängig ist.

Da ich noch nicht das Gefühl hatte, alle Einflüsse auf die Qualität von Conclusions-

Abschnitten gefunden zu haben und mit dem aktuellen Stand die Zusammenhänge noch nicht detailliert beschreiben konnte, wandte ich weitere Techniken an, um meine theoretische Sensitivität zu steigern:

Ich versuchte, mit meinem bisher angesammelten Wissen und meiner theoretischen Sensitivität ein positiv Beispiel eines Conclusions-Abschnitts zu formulieren und ein negatives Beispiel. Die von mir formulierten Beispiele spiegelten stark meinen zuvor gewonnen Erkenntnissen bezüglich der Faktoren wieder.

Aufgrund mangelnder Ansätze las ich erneut das Kapitel zum Codieren von Strauss und Corbin und versuchte es erneut mit Microcoding. Dies war nicht sehr erfolgreich. Es erschien mir so, als ob die dargestellten Techniken nur dann hilfreich wären, wenn man das Thema des Textes besser und umfassender verstehen wollte. Für meine Forschung war aber nicht der besprochene Inhalt relevant, sondern eher die Präsentation des Textes selbst. In den angeführten Beispielen von Strauss und Corbin war die Präsentation einer Aussage lediglich ein weiterer Hinweis auf den Inhalt bzw. dessen Bedeutung selber.

Zusammenfassung und Reflexion Mir ist aufgefallen, dass ich die GTM-Techniken von Strauss Corbin nicht systematisch angewandt habe und häufiger meine Perspektive während der Analyse hätte wechseln sollen. Bevor ich mich auf eine neue Perspektive einließ, schien es mir wichtig, die erste Perspektive erst vollständig zu durchdringen, vor allem weil meine erste Perspektive die strukturell inhaltliche war, die Vertrauteste für mich. Im Nachhinein ist mir bewusst, dass ein umfassendes Verständnis einer Perspektive viel Zeit erfordert, aber nicht perfekt sein muss, um nützlich zu sein. Zudem erkenne ich, dass das Betrachten aus einer anderen Perspektive auch für das generelle Verständnis der initial gewählten Perspektive hilfreich gewesen wäre.

Ein Beispiel dafür ist mein Versuch, Microcoding anzuwenden. Dadurch dass die Beispiele von Strauss Corbin sich hauptsächlich damit beschäftigen den Inhalt bzw. das Thema der Daten näher zu ergründen, wandte ich Microcoding hauptsächlich auf die rein inhaltliche Ebene der verschiedenen Conclusions-Abschnitte an. Damit kam ich schnell zu dem Schluss, dass Microcoding für meine Forschung nutzlos sei, da die Informationen, die ich mit der inhaltlichen Perspektive und Microcoding in Erfahrung brachte, nicht Gegenstand meiner Forschung waren.

Während ich das Vergleichen von Extremen anwandte, wurde mir außerdem bewusst, dass eine persönliche Eigenschaft von mir meine Arbeit stark beeinflusst hatte. Meine generelle Arbeitsweise folgt dem Konzept „Erst die Arbeit, dann das Vergnügen“ Um dies einzuhalten, führe ich zwei Listen. Eine Liste enthält Aufgaben, die erledigt werden müssen, aber keine Freude bereiten, die andere Liste enthält Aufgaben, die Freude bereiten. Wann immer mir eine vergnügliche Aufgabe einfällt, halte ich inne und notiere mir diese. Danach fahre ich mit der aktuellen Aufgabe fort und erledige noch eine weitere Pflichtaufgabe, die mir weniger Freude bereitet, bevor ich mich einer Aufgabe zuwende, die mir Freude bereitet. Normalerweise komme ich damit sehr gut zurecht und effektiv voran. Für einen Prozess der kreativen Theorienfindung ist diese Arbeitsweise eher hinderlich, denn spontane Vermutungen und Ideen direkt zu überprüfen waren für mich Aufgaben, die Freude bereiteten und wurden daher nur als Memos für später festgehalten. Im Laufe meiner Forschung bemerkte ich, dass sobald ich Zeit hatte, um einer Vermutung nachzugehen, die mir eigent-

lich Freude machen sollte, der Spaß verschwunden war und die Vermutung schwer zu greifen war. Somit vernachlässigte ich einige Vermutungen zu unüberlegt. Nachdem mir dieses Verhalten bewusst war, habe ich mich von Zeit zu Zeit dabei ertappt, dass ich diese Vorgehensweise fortsetzen wollte, konnte mich aber davon abhalten. Ich vermute, dass ich durch diese Arbeitsweise oft den Faden einer neuen Hypothese abgeschnitten habe, bevor ich ihn weiterverfolgen konnte.

5.2 Überprüfung meiner gefundenen Qualitätsfaktoren

Da aus dem letzten Teil meiner qualitativen Analyse, dem Vergleichen von Extremen, die Qualitäts-Faktoren für Conclusions-Abschnitte entstanden sind und dieser Vergleich auf meiner subjektiven Einschätzung beruht, habe ich im Anschluss eine informelle Studie durchgeführt. Diese Studie wird im Folgenden beschrieben.

Studienaufbau: Die informelle Studie sollte meine auf subjektiven Einschätzung basierende Forschung überprüfen, weshalb ich zu Beginn überlegte, mich an der *Q Methodology* [11] zu orientieren, mit Hilfe derer die Subjektivität von Aussagen untersucht werden kann. Ähnlich wie bei der Q Methodology könnte ich mit der informellen Studie meine gefundenen Qualitäts-Faktoren bezüglich deren Subjektivität beurteilen lassen.

Als Zielgruppe kamen für diese Studie alle Forschenden des SWE in Frage, da diese regelmäßig wissenschaftliche SWE-Artikel lesen. Aufgrund der begrenzten Zeit, hat mir Prof. Dr. Prechelt dabei geholfen, die Teilnehmenden zu rekrutieren. Da ich mit der Q Methodology nicht vertraut war und meine Studie aus Zeitgründen und Verfügbarkeiten möglicher Teilnehmenden möglichst schlank in der Erstellung und in der Durchführung sein sollte, entschied ich mich letzlich für eine überwachte Umfrage, in der Studienteilnehmende ihre Antworten jeweils begründen sollten und die daher eher die Form eines strukturierten Interviews hatte.

Die Studie bestand aus zwei Teilen. Im ersten Teil sollten komplette Conclusions-Abschnitte bewertet werden, im zweiten Teil einzelne Schlussfolgerungs-Sätze. Der erste Teil beinhaltete komplette, aber kurze Conclusion-Abschnitte, die ich so auswählte, dass mindestens ein schlechter und eher guter Abschnitt vorhanden waren. Außerdem achtete ich darauf, dass die ausgewählten Abschnitte verschiedene meiner gefundenen Faktoren enthielten, die sich nicht innerhalb eines Satzes darstellen ließen, wie eine offensichtliche Strukturierung des Artikels. Im zweiten Teil sollten immer zwei verschiedene Ausprägungen eines Satzes gegeneinander bewertet werden. Dabei war einer der Sätze ein Original aus meiner codierten Datenmenge und der andere eine Abwandlung davon, die die gegensätzliche Ausprägung eines Qualitätsfaktors darstellte. Beim Erstellen der Sätze, habe ich darauf geachtet, nur das Merkmal des Faktors zu verändern. Gleich bleiben sollte der Sinn, die Grammatik und die generelle Ausdrucksweise, solange diese nicht Teil des Merkmals waren. Ich habe die Sätze bewusst so angeordnet, dass nicht alle meiner Meinung nach besseren Sätze immer die erste Version bzw. zweite Version des Satzes waren, um vorzubeugen, dass Teilnehmende sich danach richten könnten. Die direkte Ausprägungen einzelner Beispielsätze miteinander zu vergleichen, schien mir passend. So konnte ich Faktoren isoliert darstellen und überprüfen, ob andere Forschende meiner Einschätzung zu-

stimmten, ohne dass ich diese preisgeben musste und die teilnehmenden Forschenden damit beeinflusste. Meine anfängliche Sorge, dass das Gegenüberstellen von gleichen Sätzen unterschiedlicher Ausprägung zu offensichtlich wäre, verflüchtigte sich nach den ersten Durchführungen der Studie. Bereits diese enthielten Einschätzungen, die gegensätzlich zu meinen Eigenen waren.

Initial sah ich für beide Teile der Studie eine vierstellige Likert-Skala vor, die von „gut“, „eher gut“, „eher schlecht“ bis zu „schlecht“ ging und mit derer die Teilnehmenden ihren Eindruck zu jedem einzelnen Satz ausdrücken sollten.

Um die Studie möglichst leichtgewichtig zu halten, wählte ich für die Satzbeispiele, wie für die Conclusion-Abschnittsbeispiele, eher kurze Sätze aus.

Um mein Studiendesign zu überprüfen und eine belastbare Einschätzung der benötigten Zeit zu erhalten, habe ich einen Pilottest durchgeführt. Die Studie sollte maximal 30 Minuten dauern. Da es mir an Teilnehmenden aus dem SWE für die Pilotierung mangelte, wurde der Pilottest mit einer ehemals forschenden Person aus dem Bereich der Geisteswissenschaft durchgeführt. Es stellte sich heraus, dass eines meiner Conclusions-Abschnittsbeispiele zu lang und ermüdend war. Dieser Effekt, war genau der, den ich nachweisen wollte. Trotzdem entschloss mich dazu das Beispiel zu entfernen, da ich nicht gleich zu Beginn die Studienteilnehmenden ermüden wollte und der Text außerdem mehrere Minuten Lesezeit benötigte. Die Zusammensetzung meiner Beispielsätze veränderte ich nicht. Ich erweiterte sie aber, da dies zeitlich möglich war und ich so je Faktor mehrere Beispiele hatte. Damit verringerte ich die Gefahr, dass Beispiele unbrauchbar wurden durch mögliche von mir nicht bemerkte Unterschiede. In diesem Pilottest funktionierte die Nutzung der initialen Skala ausreichend gut. Nach den ersten zwei Durchführungen der Studie, stellte sich diese Skala als unhandlich heraus. Daher änderte ich diese in ein intuitiveres Format um. Es sollte nicht mehr jeder Satz einzeln bewertet wurde, sondern es sollte entschieden werden welcher der Sätze besser sei und wie viel besser. Die neue Skala beschrieb den Unterschied zwischen Sätzen bzw. Conclusions-Abschnitten mit „viel besser“, „besser“, „eher besser“ oder, im Fall von keinem wahrgenommenen Unterschied in der Qualität, mit „gleich“.

Damit die Studie meine Ergebnisse unterstützen würde, sollten die Einschätzungen der Teilnehmenden meinen Eigenen entsprechen. Dafür würde es ausreichen, wenn fast alle der Teilnehmenden zustimmen würde und es kaum gegensätzliche Bewertungen gäbe. Von gegensätzlichen Meinungen versprach ich mir auf Perspektiven oder Gegebenheiten aufmerksam gemacht zu werden, die ich selbst übersehen hatte. Um dies auch zu erreichen, sollten Teilnehmende nicht nur auf einer Skala bewerten, sondern ihre Bewertungen auch begründen.

Die Einschätzungen und Begründungen haben ich auf einem vorbereiteten Dokument festgehalten (Bsp.: siehe C.3). Um Begründungen zu Einschätzungen zu erhalten, habe ich Teilnehmende aufgefordert, laut zu denken. Damit ich meine eigene Meinung nicht unwillentlich kommunizierte, wenn ich Nachfragen zu Begründungen hatte, habe ich mir in der Vorbereitung der Studie, notiert worauf ich hinweisen dürfte. Während der überwachten Umfrage habe ich mich daher auf Folgendes beschränkt: Einfache neutral formulierte Nachfragen, wie: „Warum gefällt dir Version A besser als B?“, „Woher kommt deine Abneigung zu B?“ oder „Welcher Teil des Satzes stört dich besonders? Warum?“, Hinweise, die den Teilnehmenden helfen soll-

ten Einschätzungen zu begründen bzgl., möglicher Annahmen bzw. Erwartungen der Teilnehmenden („Liegt deiner Beurteilung evtl. eine Annahme zu Grunde? Welche?“) und Hinweise auf verschiedene mögliche Betrachtungsebenen wie Grammatik und empfundene Lesbarkeit.

Die vollständige Studie mit Notizen für meine Studieneinleitung und allen genutzten Beispielen und meiner Vorlage zum Festhalten der Ergebnisse ist im Anhang zu finden (Anhang C).

Studiendurchführung: Jeder Durchlauf meiner Studie hat pro teilnehmender Person ca. 30 Minuten gedauert und begann mit einer kurzen Einleitung, in der ich das Thema meiner Masterarbeit erklärte, den Zweck und das Vorgehen dieser informellen Studie. Besonders wichtig dabei war es, auf meine Eingangsannahme hinzuweisen, dass ich innerhalb von Conclusions-Abschnitten Schlussfolgerungen erwarte und Conclusions-Abschnitte auch mit dieser Annahme bewertet werden sollten.

Um die überwachte Umfrage mehr wie ein strukturiertes Interview erscheinen zu lassen, habe ich die Teilnehmenden ihre Antworten nicht selbst ankreuzen lassen, sondern mir diese mündlich mitteilen lassen und per Hand notiert. Damit hatten die Teilnehmenden nur ein Antwortmedium zu beachten (das mündliche). Diese Vorgehensweise hatte den zusätzlichen Vorteil, dass ich Begründungen immer direkt neben den auf der Skala festgehaltenen Bewertungen notieren konnte, was die Auswertung später einfacher machte. Nachteilig daran war, dass manche Teilnehmende ihre eigene Bewertung vergaßen und ich trotz sofortiger Nachfrage zu Begründungen ihnen diese nochmal nennen musste. Manchmal vergaßen Teilnehmende auch, zu bewerten wie viel besser sie Artikel im Vergleich zueinander fanden.

5.2.1 Ergebnisse der Studie

Insgesamt hatte ich acht Teilnehmende. Es handelt sich bei allen Teilnehmenden um Forschende, die aktuell im Bereich des SWE forschen, mit unterschiedlichen jahren an Erfahrung in diesem Bereich: von vor einem Jahr in die Wissenschaft Eingestiegene bis zu Personen mit langjähriger Professur.

Meine Annahme, über das Vorhandensein von Schlussfolgerungen innerhalb von Conclusion-Abschnitten, wurde von allen Teilnehmern als einleuchtend angenommen. Sechs der Teilnehmer stimmten dieser Annahme unaufgefordert zu.

Ergebnisse der kompletten Conclusions-Abschnitte: Meiner Beurteilung nach war das zweite Beispiel das Beste, folgen tat das erste Beispiel und das schlechteste Beispiel war das dritte. Die Beispiele sind im Anhang (siehe Unterabschnitt C.1) zu finden. Das erste Beispiel beurteile ich als Bestes, der drei Conclusion-Abschnitte, da es konkrete Empfehlungen und damit Schlussfolgerungen liefert und diese auch deutlich als solche kennzeichnet. Und zwar durch die Verwendung der ersten Person, was die Lesbarkeit des Artikels stark erhöht. Die anderen beiden Artikel geben das eigentliche Wissen, was durch jeweilige Studien gewonnen wurde nicht weiter. Es wird lediglich beschrieben, dass etwas gefunden wurde und was. Während dies in der ersten Studie etwas ausführlicher getan wird, wird dies in der letzten Studie nur sehr knapp beschrieben und lässt mich als Lesende mit keiner nutzbaren Information

zurück. Alle drei Artikel haben gemeinsam, dass die Sätze eher einfach formuliert sind.

Sieben der Acht Teilnehmenden haben die Artikel so angeordnet, wie ich es erwartet hatte. Sechs der Teilnehmenden, die nach meiner Erwartung ordneten, beurteilten die Conclusions-Abschnitte hauptsächlich nach Inhalt, vor allem bezüglich, Schlussfolgerungen, entsprechend meiner in der Einleitung erläuterten Annahme. Das Vorhandensein von Schlussfolgerungen war das wichtigste Beurteilungskriterium für alle der sieben Teilnehmer, möglicherweise aber auch nur, weil ich dies in meiner Einleitung angegeben hatte. Aber nachfolgend zu dem Informationsgehalt bezüglich der Schlussfolgerungen war es den Teilnehmern auch wichtig, eine kurze Zusammenfassung über die Studie zu lesen. Dadurch wurde das dritte Beispiel immer als das schlechteste eingestuft, da die Teilnehmenden nach dem Lesen des Abschnitts fast genauso viel wussten, wie zuvor.

Folgende weitere Anmerkungen für den ersten Artikel wurden gegeben:

- Die Eindeutigkeit, mit der die Schlussfolgerungen präsentiert wurden, wurde von vier Teilnehmenden als positiv hervorgehoben. Alle Teilnehmenden benannten in diesem Kontext die Empfehlungs-Form, in der Schlussfolgerungen gegeben wurden. Eine Person nannte zusätzlich die Aufzählungs-Form in der die Schlussfolgerungen gegeben wurden.

- Zwei Teilnehmenden fiel auf, dass der Conclusions-Abschnitt keine Resultate enthielt. Daher war unklar, wie die Autoren zu den Schlussfolgerungen gekommen waren. Nur einer von ihnen fiel dies in einem negativen Sinn auf.

- Zwei Teilnehmende fanden es gut, dass der erste Conclusions-Abschnitt eine kurze Zusammenfassung der Studie enthielt, da diese ihnen das Verständnis der nachfolgenden Schlussfolgerungen vereinfachte. Eine dieser Personen schweifte allerdings beim Bewerten der Artikel ab und beurteilte nach inhaltlicher Verständlichkeit, was nicht Teil der Aufgabe war.

Das erste Beispiel eines Conclusions-Abschnitt erschien den Teilnehmenden immer besser als das dritte Beispiel, obwohl beiden Beispielen Schlussfolgerungen fehlen. Das erste Beispiel hat mehr Textvolumen und enthält daher mehr Informationen als das dritte Beispiel. Von fünf Teilnehmenden der Studie wurde das so ausgedrückt, dass der dritte Artikel der schlechteste ist, weil dieser gar keine Informationen über die Studie enthält. Nach dem Lesen des Artikels sei weder klar, was getan wurde, was herausgefunden wurde, noch wozu die nicht gegebenen Schlussfolgerungen nützlich seien. Obwohl das Benennen von Schlussfolgerungen das dritte Beispiel unweigerlich verbessert hätten, merkte eine teilnehmende Person an, dass sie zwar gerne konkrete Schlussfolgerungen sehen würde, aber auf gar keinen Fall die gesamte Liste der 28 gefundenen besten Praktiken im Conclusions-Abschnitt.

Positiv am ersten Beispiel wurde von drei Teilnehmenden hervorgehoben, dass dieses eine gute Beschreibung der Studie enthielt und damit die Glaubwürdigkeit der Studienerkenntnisse erhöhe. Das erste Beispiel enthält damit zwar noch immer die Schlussfolgerungen vor, aber erkläre das Zustandekommen und den Nutzen der Schlussfolgerungen zumindest ausreichend, so dass Lesende in der Lage wären, zu beurteilen, ob sie den Artikel lesen wollten oder nicht. Eine teilnehmende Person

merkte zum ersten Beispiel an, dass dieses dazu tendiere, nur Resultate und keine Schlussfolgerungen zu nennen und dies nicht ausreichend sei.

Als abschließende Anmerkungen zu allen drei Artikeln wurde von einer Person erwähnt, dass alle Conclusions-Abschnitte sich selbst auf die Schulter klopfen würden und von drei Teilnehmenden wurde angemerkt, dass keines der Beispiele besonders schlimm gewesen sei und sie Schlimmeres gewohnt wären.

Die teilnehmende Person, die die Artikel anders als erwartet sortierte, empfand alle Abschnitte als schlecht. Bei allen Artikeln fehlte ihr die Beantwortung der Frage, was die Erkenntnissse konkret bedeuteten, also was die Auswirkungen der Erkenntnisse waren. Sie sortierte den zweiten Artikel als am wenigsten schlecht ein, weil der als einziger konkrete Ratschläge enthielt, auch wenn bei diesen der größere Kontext fehlte. Als den zweitschlechtesten Conclusions-Abschnitt ordnete sie Beispiel 3 ein. Dieser enthalte im Gegensatz zu Beispiel 1 immerhin einen Satz, der den Beitrag des Artikels beschreibt. Die Person hatte eine sehr konkrete Erwartung an Conclusions-Abschnitte. Innerhalb von Conclusions-Abschnitten sollten neben Resultat nahen Schlussfolgerungen auch Schlussfolgerungen einer höheren Abstraktionsebene enthalten sein, z.B. in Form von Ratschlägen, deren Nutzen erklärt wird.

Ergebnisse der einzelnen Satz-Beispiele: Im Rahmen der informellen Studie habe ich meine gefundenen Qualitäts-Faktoren überprüfen wollen. Negative Qualitätsfaktoren sind Faktoren, deren Auftreten innerhalb eines Conclusions-Abschnitts diesen eher verschlechtern. Für positive Faktoren gilt das Gegenteil. Eine vollständige und ausführlichere Liste aller Faktoren mit allen Ergebnissen ist in Unterunterabschnitt 5.3.2 zu finden.

1. **Vorenthaltene Informationen** sind ein negativer Qualitäts-Faktor, weil die Erkenntnisse des Artikels zu unkonkret bleiben. Dieser Faktor beschreibt Schlussfolgerungen, die keine konkreten Informationen zur Schlussfolgerung benennen. So wird Lesenden zwar bewusst, dass die Autoren woanders im Artikel Schlussfolgerungen gezogen haben und möglicherweise sogar auch welche Schlussfolgerungen, aber die Schlussfolgerungen werden im Conclusions-Abschnitt vorenthalten.

2. **Eindeutige Signale für Schlussfolgerungen** signalisieren Lesenden in einer offensichtlichen Weise, dass eine Schlussfolgerung präsentiert wird, und sind ein positiver Qualitätsfaktor.

3. **Füllsätze** sind Sätze, die Autoren nutzen, um die Bedeutung ihrer Schlussfolgerungen zu betonen. Die Aussage dieser Sätze ist dabei oft schwer zu überprüfen und es werden keine weiteren Informationen gegeben. Solche Sätze treten häufig in Kombination mit schwach formulierten Schlussfolgerungen auf. Füllsätze zählen daher zu den negativen Qualitätsfaktoren.

4. **Squishy wording** zählt zu den negativen Faktoren und beschreibt Schlussfolgerungssätze, deren Inhalt durch die Verwendung einzelner Wörter, wie z.B. „vielleicht“, weniger bestimmt bzw. schwammiger wird. Dies ist ein negativer Qualitäts-Faktor.

5. **Komplexer Satzbau** ist ein negativer Faktor und beschreibt Sätze, die durch ihre Komplexheit schwerer zu verstehen sind.

Die Faktoren 5. Komplexer Satzbau und 4. Squishy wording können mit den Ergebnissen eindeutig bestätigt werden. Keine teilnehmende Person bewertete die gegebenen Beispiele gegensätzlich zu meiner Erwartung. Bei allen Beispielen der beiden Faktoren bewertete die Mehrheit die Sätze als besser. Auch die gegebenen Begründungen stimmen in der Mehrheit mit meinen Erwartungen überein.

Der Faktor 3. Füllsätze wurde kaum so bewertet und begründet, wie ich es erwartet hatte. Die Begründungen zur Bewertung betreffender Beispiele hatte die größte Variabilität. Bei Beispiel *HatNovBal-2-3* beurteilten manche Teilnehmenden die beabsichtigten Faktoren anders als erwartet: Ein verlängerter Absatz durch einen Füllsatz soll negativ wirken und eindeutige Signalisierung der Schlussfolgerungen durch die Nutzung der ersten Person soll positiv wirken. Zwei der Teilnehmenden nehmen den Unterschied in der passiven und aktiven Schreibweise (erste Person bzw. neutrale Formulierung) als eher negativ auf, wobei eine der teilnehmenden Person dies als persönliche Vorliebe einordnet und die Länge des Absatzes für beide Varianten vertretbar findet. Diese teilnehmende Person stuft daher beide Sätze als neutral ein. Nur zwei Teilnehmende begründen ähnlich meiner Erwartungen, dass ihnen Version A besser gefalle, da dieser konkreter und vor allem kürzer erscheine. Andere Teilnehmende konnten dies genauer Beschreiben und gaben an, dass ihnen in Variante B Informationen fehlten und unklar bleibe, wofür und für wen die Empfehlungen seien. Diese Information war im Füllsatz angesehen, der entgegen meiner Wahrnehmung eine relevante Information enthielt. Insgesamt sehe ich diesen Faktor damit als unpassend an, da der Füllsatz des zweiten Beispiels für mindestens drei der Teilnehmenden wichtige Informationen enthielt, die ich als so selbstverständlich angesehen hatte, dass ich sie als nicht nennenswert eingestuft hatte.

Zu den anderen Faktoren gibt es keine so eindeutigen Ergebnisse. Für den Faktor 1. Vorenthaltene Informationen widersprechen sich die Ergebnisse der beiden Beispiele. Wärhrend das Beispiel *MarSen-1* von allen Teilnehmenden passend zu meiner Erwartung beantwortet wurde, sind die Antworten von zwei Teilnehmenden zum Beispiel *BalRal-1* gegensätzlich meiner Erwartung. Den gegebenen Begründungen nach liegt dies daran, dass *BalRal-1* in Variante A einen komplizierteren Satzbau hat. Damit unterscheiden sich Variante A und B nicht nur im Informationsgehalt, wie von mir beabsichtigt, sondern auch in der Komplexität des Satzbaus. Dass die Komplexität des Satzbaus hier als wichtigster Faktor wahrgenommen wird, bestätigt Faktor 5. Komplexer Satzbau. Trotzdem bestätigt die Mehrheit der Teilnehmenden Faktor 1. Vorenthaltene Informationen in dem Sinne, dass Schlussfolgerungen möglichst konkret gegeben werden sollten. Teilnehmende haben aber nicht hervorgehoben, dass das Vorhandensein konkreter Verweise in den Hauptteil der Studie positiv sind. Solche Verweise wurden als negativ beschrieben, hatten dabei aber weniger Einfluss auf die Gesamtbeurteilung des Satzes als die Konkretheit von Schlussfolgerungen.

Faktor 2. Eindeutige Signale für Schlussfolgerungen hat gemischte Ergebnisse. Bei einem Beispiel stimmt die Mehrheit der Teilnehmenden meinen Erwartungen zu, bei dem anderen Beispiel (*PreRob-2*) nicht. Beim Beispiel *PreRob-2* ist mir im Nachhinein aufgefallen, dass beide Varianten Hinweise dazu enthalten, dass sie Schlussfolgerungen sind:

- Variante A (original) „In general, we advise against pre-training a new model from scratch as it is extremely resource intensive, for mixed results“ [65].
- Variante B „Pre-training a new model from scratch is extremly resource intensive, for mixed results.“ [65]

Variante A ist offensichtlich eine Schlussfolgerung, durch das Signalwort „In general“ und der Form einer Empfehlung „we advise“. Variante B ist eindeutig eine Schlussfolgerung, durch den letzten Teilsatz. Das habe ich übersehen. Das Beispiel ist daher ungültig. Auch die Begründungen zu diesen Beispielen und anderen Beispielen anderer Faktoren sind alle weder passend noch unpassend bezüglich des Nutzens eindeutiger Signale und halten sich die Waage. Aus Begründungen geht hervor, dass die Relevanz für eindeutige Signale für Schlussfolgerungen je nach Verortung des Satzes variiert. So ist ein eindeutiges Signal für eine Schlussfolgerung direkt zu Beginn des Abschnitts wichtig, aber in der Mitte oder zum Ende des Abschnitts anscheinend nicht. So gab eine teilnehmende Person, die die passive Schreibweise persönlich bevorzugt, an, dass sie für eine einleitende Schlussfolgerung ein eindeutiges Signal in Form einer persönlichen Schreibweise bevorzugen würde.

Die vollständigen und detaillierteren Ergebnisse sind im Anhang C.4 zu finden.

5.3 Erkenntnisse: Was beeinflusst die Qualität von Conclusions-Abschnitten?

In diesem Abschnitt beschreibe ich vor allem meine Schlussfolgerungen, die für die Bewertung der Qualität der Conclusions-Abschnitte im Bezug auf Schlussfolgerungen relevant sind. Dieser Abschnitt enthält Beispiele und Details zu meinen Schlussfolgerungen.

Folgend eine Auflistung allgemeiner Schlussfolgerungen bezüglich der Qualität und des Aufbaus von Conclusions Abschnitten:

- Durch die Ergebnissen aus meiner informellen Studie bin ich darauf aufmerksam geworden, dass ich Conclusions-Abschnitte während meiner Forschung zu eng betrachtet habe. Mit meiner eingangs getroffenen Annahme, dass Schlussfolgerungen der wichtigste Teil von Conclusions-Abschnitten sind, habe ich mich nur noch auf Schlussfolgerungen fokussiert. Neben Schlussfolgerungen ist aber zumindest einer teilnehmende Person der informellen Studie auch der Ausblick und die Vision der Forschung wichtig, da die Forschung ohne solches schlechter in das Forschungsfeld und die Realität eingeordnet werden kann. Dies erklärt z.B. den Zweck von fhope-Sätzen, die, entgegen meiner bisherigen Annahme, keine unsichere Art der Ausdrucksweise sind, sondern eine Vision beschreiben, die die Forschung in einen Kontext setzt. Ich befürchte, dass alle von mir gefundenen Faktoren zur Qualität von Conclusions-abschnitten sich nur auf Schlussfolgerungen beziehen und alle anderen möglicherweise relevanten Merkmale eines guten Conclusions-Abschnitts vernachlässigen.
- Die inhaltliche Struktur von Conclusions-Abschnitten sagt wenig über die Qualität des Abschnitts aus, sie ist hilfreich als Grundlage zur Erforschung anderer Merkmale. Durch diese Codes konnte ich ein besseres Textverständnis für Conclusions-Abschnitte entwickeln, wodurch ich in der Lage war, auch andere Betrachtungsebenen einzunehmen.

- Die Struktur von Conclusions-Abschnitten ähnelt der eines Abstracts.
- Die Qualität der Conclusions-Abschnitten, hängt in erster Linie vom Verhältnis von Schlussfolgerungs bezogenem Text zu Studienrekapitulierenden Text und der Lesbarkeit und Konkretheit der Schlussfolgerungen ab. Beide können durch unterschiedliche Faktoren beeinflusst werden.

Insgesamt hatte ich am Ende meiner Arbeit 48 Artikel codiert und analysiert, von denen ich zehn als gute Conclusions-Abschnitte einschätzte, da sie ihre Schlussfolgerungen geordnet und klar vermittelt haben und 21 als schlecht, da diese kaum Schlussfolgerungen präsentierten.

Folgend möchte ich komplette Beispiele zur Veranschaulichungen von Qualitätsmerkmalen in Conclusions-Abschnitten geben:

Beim Lesen des Textes in Abbildung 2 wird deutlich, dass der Hauptzweck des Abschnitts darin besteht, neue Erkenntnisse aus der Studie zu vermitteln. Das Wissen ist verständlich aufbereitet, d.h., es ist gut lesbar und die Informationen werden konkret im jeweiligen Satz gegeben. Die Lesbarkeit wird in diesem Text vor allem durch folgende Punkte unterstützt:

- Das Verhältnis von Text bzgl. Schlussfolgerungen (Grün und Türkis) zu Text mit Beschreibung der Studie (Blau Töne) ist etwa eins zu drei. Die Einleitung des Abschnitts ist nicht zu lang. Damit ist dieser Conclusions-Abschnitt eindeutig darauf fokussiert, die Erkenntnisse zu präsentieren.
- Alle Schlussfolgerungen werden eindeutig als solche formuliert, indem sie als Empfehlungen formuliert werden, z.B. „We recommend selecting ILA according to particular purpose“ [48]. Diese Empfehlungen wirken durch die verwendete erste Person, als ob die Autoren dahinterstehen.
- Die Art, wie der Text strukturiert ist, verbessert die Lesbarkeit. Durch die Struktur des Textes können Lesende Schlussfolgerungen einfacher erkennen. In diesem Text wird dies durch eine Art Auflistung erzeugt, die Empfehlungen auch klar als solche ankündigt.
- Es werden einfache, kurze, Sätze verwendet, die ebenfalls die Lesbarkeit verbessern.
- Die Konkretheit wird hier vor allem durch das Vorhandensein aller Informationen zu den Schlussfolgerungen und durch das Vorstellen der Schlussfolgerungen selbst erreicht. Der einzige Verweis im Text ist genau und verweist auf weitere Detaillinformationen und enthält die wichtigsten Schlussfolgerungen nicht vor.

Der Conclusions-Abschnitt in Abbildung 3 ist schwieriger zu lesen als das vorangegangene Beispiel und vermittelt Schlussfolgerungen deutlich weniger eindeutig. Der Abschnitt ist auch weniger verständlich strukturiert

- Der studienbeschreibende Teil (Blau-Töne) beschreibt die Motivation des Artikels ausführlich und nimmt dabei insgesamt fast die Hälfte des Conclusions-Abschnittes ein. Für die meisten Lesenden, die mindestens schon das Abstract

9 Conclusion

The impact of false-positive/negative defect-inducing commits on the defect prediction performance is important when evaluating defect prediction models. To reduce the number of false-positive/negative defect-inducing commits, many prior studies have proposed ILAs to detect defect-fixing commits accurately (Fischer et al. 2003a; Śliwerski et al. 2005; Bachmann and Bernstein 2009a, 2010; Bird et al. 2010; Sureka et al. 2011; Wu et al. 2011; Nguyen et al. 2012; Bissyandé et al. 2013; Le et al. 2015; Schermann et al. 2015; Sun et al. 2016, 2017a, b; Xie et al. 2019; Tu and Menzies 2020). However, challenges still exist, such as dataset inconsistency and small comparisons. Our work is the first large-scale study to evaluate the ILAs on the same experimental settings. In addition, we summarized the prior ILAs as our related work through our systematic literature study.

In the following, we summarize the main recommendations. We recommend selecting ILAs according to the particular purpose.

Recommendation 1: For researchers and practitioners who need to evaluate defect prediction models in the low-quality dataset scenario, we recommend using the combination of the natural language text similarity, Phantom heuristics, random forest, and support vector machine approaches. Our experiments in RQ2 show that the studied ILAs prevented the defect prediction model from being affected by missing defect-fixing commits. In particular, the combination of the natural language text similarity, Phantom heuristics, random forest, and support vector machine approaches achieved the most similar defect prediction performance with the ground-truth defect prediction performance across all the studied ILAs. If researchers and practitioners use this combination, they would not need to worry about the impact of missing defect-fixing commits to defect prediction performance.

Recommendation 2: For researchers and practitioners who need defect-fixing commits without false-positive defect-fixing commits, we recommend using the random forest approach. Our experiments in RQ1 have shown that the random forest approach achieved the highest precision. Hence, using this ILA rarely induces false-positive defect-fixing commits while reducing missing defect-fixing commits.

Recommendation 3: For researchers and practitioners who need defect datasets without missing defect-fixing commits, we recommend using the time filtering approach or the natural language text similarity approach. Our experiments in RQ1 have shown that the time filtering approach and the natural language text similarity approach performed the highest TP rate compared with the other accurate ILAs in different projects. Hence, using the time filtering approach or the natural language text similarity approach reduces the number of missing defect-fixing commits. In addition, these ILAs are very simple. Note that when using these approaches, researchers and practitioners need to consider false-positive defect-fixing commits.

Recommendation 4: Considering the dates of the commit and the issue report help exclude noise of defect-fixing commits. Our manual inspection discussed in Section 7.5 shows that our basic restriction can exclude defect-fixing commits that do not fix source code. Hence, using our basic restriction that uses the dates excludes such noise of defect-fixing commits for defect prediction.

Abbildung 2: Beispiel für einen guten Conclusions-Abschnitt. Erklärung der Farben: Dunkelblau-Hintergrundinformationen; Hellblau-Methode; Schwarz-Überleitung; Grün-Schlussfolgerung; Türkis-Resultat das Schlussfolgerungen erläutert; dunkleres Türkis-Einschränkungen

8 | CONCLUSIONS AND FUTURE WORK

The Augmented Web is very powerful and an interesting way to allow end-users to customize the way in which they surface the Web. However, it comes with particular challenges, given that both scenarios of production and testing are very diverse because of the big crowd of end-users and Web sites. This is even exacerbated because the adapted content on the client-side is not always the same for every user, given the features of modern Web applications, which depend on platforms and sessions. With this in mind, each augmentation session constitutes a production ambient in itself, and to allow the generation of public and automatized reports when the augmenter fails is determining. We strongly believe that a key point in the adoption of augmentation technologies is the correct working of augmenters. When augmenters fail because the target Web site has been upgraded, it is necessary to notice it, test if the augmenter is compatible with the current version of the DOM and, if it is not compatible, allow the crowd of users of the augmenter to fix the references in order to continue using it correctly as soon as possible. If this process is not supported by a methodological approach accompanied with specific tools, end-users may be frustrated and uninstall the augmenter before it is maintained by its developer.

In this paper, we proposed a method for developing augmenters without hard-coded references to Web sites' DOM and for generating the augmenter source code that includes a visual tool for end-users to maintain broken DEOI references. Our approach has been evaluated in comparison with traditional maintenance of augmenters. Considering a two days time lapse, the results show the convenience of our approach and the feasibility of maintaining augmenters by end-users.

Even though hard-coding data is a bad smell [37], as we also described for augmenters before, hard-coding references to augmented websites is currently an Achilles heel for augmenters. Our approach proposes to breathe life into these references by treating them as dedicated objects rather than data clumps. We strongly believe that our approach, where end-users are allowed to share and maintain collectively these DEOI references, facilitates the evolution of the augmenters which we have shown in our evaluation.

Adaptive development and maintenance have been highlighted in the context of user-centric software artefacts [38]. According with our study, userscripts are actually user-centric software artefacts because their developers are usually actual users of them, that share these artefacts in the community, reaching more users who contribute with the evolution of the augmenters by commenting on these repositories.

The kind of study presented in this paper is related to existing ones in mash-up communities, specifically for Yahoo! Pipes [39]. In that work, Stolee et. al discovered, that in these communities, configurable artifacts are desired. Also they conclude "here is a need for better end-user programmer support in several stages of the software life-cycle, including development, maintenance, search, and program understanding", which we have demonstrated that it is also important in the case of userscripts.

In the future, we plan to improve the support for adaptive or perfective maintenance of augmenters in order to support backward compatibility in those cases in which developers have written source code after generating the augmenter with our tool. Finally, also the tool supporting end-users for maintaining DEOI references may be improved in several ways to enhance usability, something that should be measured as well in the future.

Abbildung 3: Beispiel für einen schlechten Conclusions-Abschnitt. Erklärung der Farben: Dunkelblau-Hintergrundinformationen; Hellblau-Methode; Neonblau-Resultate; Schwarz-Überleitung; Grün-Schlussfolgerung; Dunkelgrün-verwandte Arbeiten; Gelb-Ausblick

9 CONCLUSION

In this article, we discuss our experience of co-creation, as means to IA collaboration, gained in three collaboration setups in the area of SE research. Throughout this experience, we have observed a set of 28 recurring best practices and issues to avoid, which we provide as the reported patterns and anti-patterns. We exemplify the patterns and anti-patterns using three different IA collaboration projects. Such exemplified insights into recurring patterns of successes and failures can positively contribute to other IA collaboration projects in SE.

Abbildung 4: Beispiel für einen schlechten Conclusions-Abschnitt. Erklärung der Farben: Blau-Studienziel; Neonblau-Resultat; Grün-Schlussfolgerung; Gelb-Ausblick

gelesen haben, wenn nicht sogar den ganzen Artikel ist die Motivation mit größter Wahrscheinlichkeit eine pure Wiederholung und nicht informativ. Lesende werden durch den langen Einführungsteil (OR) eher ermüdet als mit Informationen konfrontiert, die zum Verständnis von Schlussfolgerungen beitragen.

- Der Satzbau ist eher komplex, durch Verschachtelungen und unbestimmte Nomen, wie „it“. Dadurch ist der Textfluss des Conclusions-Abschnitts nicht optimal. Dies kann auf fehlende Überleitungen zurückzuführen sein. Wenn man davon ausgeht, dass jeder Absatz nach der Einleitung eine Schlussfolgerung darstellen sollte, wäre neben einem Überleitungssatz, der dies erklärt, auch eine stärkere Strukturierung des Textes für die Lesbarkeit hilfreich gewesen.

- Die Konkretheit leidet vor allem unter den eher unkonkreten Schlussfolgerungen, die nicht klar formuliert sind (siehe rot unterstrichen). Der erste rot markierte Satz spricht von einem Artefakt. Dieses wird weder kurz beschrieben, noch wird genauer verwiesen, wo man es lesen könnte.

Ein weiteres Beispiel für einen Conclusions-Abschnitt ist in Abbildung 4 zu sehen. Das Lesen dieses Abschnitts sollte wieder leichter fallen. Die Lesbarkeit des Abschnitts ist gut: Das Verhältnis der Einführung zu den Ergebnissen des Artikels ist angemessen und die Sätze sind einfach aufgebaut. Es wird auch häufig die erste Person verwendet. Die Autoren scheuen sich also nicht, ihre Aussagen mit sich in Verbindung zu bringen. In diesem Abschnitt wird jedoch keine der Schlussfolgerungen der Autoren geteilt. Der einzige Schlussfolgerungssatz ist beschreibend und verrät die Schlussfolgerung nicht konkret. Auch wirs kein genauer Verweis in die Studie wird gegeben, wo ein interessierter Leser nachschlagen könnte.

Dieser Abschnitt ist ein Beispiel für einen Conclusions-Abschnitt, der kaum Informationen an den Leser vermittelt und daher kein Wissen weitergibt. Zum einen weil der Abschnitt sehr kurz ist, zum anderen weil die Schlussfolgerungen, die genannt werden können, nur oberflächlich beschrieben werden. Egal wie gut die Lesbarkeit ist, solange nicht genügend Schlussfolgerungen konkret präsentiert werden, kann der Conclusions-Abschnitt nicht als gut bewertet werden. Möglicherweise haben die Autoren beim Schreiben dieser Schlussfolgerung eher das Ziel gehabt ihre Studie knapp zusammenzufassen, wie es bei einem Klappentext eines Buchs üblich.

5.3.1 Faktoren die komplette Conclusions-Abschnitte betreffen

In diesem Abschnitt wird eine Übersicht gegeben über die Qualität beeinflussende Eigenschaften die innerhalb von kompletten Conclusions-Abschnitten auftreten können und nicht nur an einzelne Sätze gebunden sind.

Das Hauptziel des Conclusions-Abschnitts ist es, die wichtigsten Schlussfolgerungen zu präsentieren. Dieser Meinung waren auch alle Teilnehmenden meiner informellen Studie. Dies ging sowohl durch das Vergleichen von kompletten Conclusions-Abschnitten hervor, bei denen alle Teilnehmenden den Conclusions-Abschnitt mit den treffendsten Schlussfolgerungen als am besten beurteilten, als auch bei einzelnen Teilnehmenden durch Anmerkungen zu den Satzbeispielen. Zwar ist hier anzumerken, dass ich vor Beginn der Bewertungen meine der Studie vorausgehende Annahme als Beurteilungskriterium nannte, dass Conclusions-Abschnitte Schlussfolgerungen enthalten müssen, allerdings widersprach keine der Teilnehmenden Personen dieser Annahme auf Nachfrage. Zudem wurde im Verlauf der Bewertungen klar, dass die Teilnehmenden in Conclusions-Abschnitten auch abseits der Studie Schlussfolgerungen erwarten.

Die folgenden Faktoren sind nach Stärke ihres Einflusses auf die Gesamtqualittät eines Conclusions-Abschnitts absteigend geordnet.

(a) **Konkretheit von Schlussfolgerungen:** Wie bereits erwähnt sind die Schlussfolgerungen der wichtigste Teil des Conclusions-Abschnitts. Ein Abschnitt ganz ohne Schlussfolgerungen kann seinen Zweck nicht erfüllen und ist damit schlecht. Schlussfolgerungen können unkonkret gehalten werden, indem sie einfach nicht gezogen werden, indem sie „angedeutet“ werden mit Resultaten und undeutlichen Formulierungen, wie im Beispiel in Abbildung 4. Das Andeuten von Schlussfolgerungen bzw. *Auslassen von Generalisierungen* ist für einen ganzen Conclusions-Abschnitt einfach erkennbar, wenn zwar viele Resultate beschrieben, aber nicht einmal unkonkrete Schlussfolgerungen gezogen werden. Ausgelassene Generalisierungen können aber auch in einzelnen Sätzen auftreten und sichtbar werden. Dies ist meist weniger offensichtlich als in ganzen Abschnitten.

(b) **Anteil an Text der Schlussfolgerungen erläutert und der die Studie beschreibt:** Conclusions-Abschnitte sollten Erkenntnisse liefern, nicht die Rechtfertigung, warum diese Erkenntnisse relevant sind und welche Methoden verwendet wurden. Beides sollte im Hauptteil der Studie und im Abstract ausreichend dargestellt worden sein. Daher sind diese Informationen allen Lesern des Conclusions-Abschnitts bereits bekannt. Eine kurze Erinnerung an verwendete Methoden kann eine Schlussfolgerung glaubwürdiger machen, eine kurze Erinnerung des Studienziels gibt mehr Kontext für die folgenden Schlussfolgerungen. Beides kann damit das Einordnen von neuem Wissen vereinfachen Das Lesen bereits bekannter Informationen, die kaum relevant sind für die von Lesenden erwarteten Schlussfolgerungen, ist ermüdend und beeinflusst daher sowohl die Lesbarkeit negativ, als auch die Motivation der Lesenden.

(c) **Strukturiertheit des Abschnitts – Listen, Absätze, roter Faden:** Dieser Faktor ist für besonders gute und leicht lesbare Abschnitte unerlässlich. Da dem Leser die Argumente der Autoren vermittelt werden sollen ist es ratsam, den Text

leicht lesbar zu gestalten. So ist es wahrscheinlicher, dass Lesende mit Sicherheit verstehen, was vermittelt werden soll. Einleuchtend verwendete Absätze für Gedankensprünge, Überleitungen und Listen für Aufzählungen mehrerer Argumente oder Schlussfolgerungen eignen sich dafür. Umgekehrt kann ein schlecht strukturierter Conclusions-Abschnitt die Lesbarkeit auch verschlechtern. Ein Beispiel für einen gut strukturierten Text ist in Abbildung 2 zu sehen. Ein schlecht strukturiertet Text ist in Abbildung 3 aufgeführt.

5.3.2 Faktoren die Schlussfolgerungs-Sätze betreffen

Dieser Abschnitt gibt eine Übersicht über Qualitäts-Faktoren, die in einzelnen Schlussfolgerungssätzen auftreten können. Wie bei den zuvor vorgestellten Faktoren für ganze Conclusions-Abschnitte beeinflussen die hier genannten Faktoren ebenfalls die Konkretheit und Lesbarkeit des Abschnitts. Der Gesamteindruck eines Conclusions-Abschnitts wird durch alle Faktoren gemeinsam beeinflusst. Kein einzelner Satz-Faktor kann die Qualität eines Conclusions-Abschnitts allein bestimmen.

Vorenthaltene Informationen: Wie stark dieser Faktor die Qualität beeinflusst, ist von zwei Eigenschaften abhängig: Die Wichtigkeit der vorenthaltenen Information und die Genauigkeit, mit der auf den Artikelteil dieser Information hingewiesen wird. Die wichtigsten Informationen, die innerhalb eines Conclusions-Abschnitts vorenthalten werden können, sind Schlussfolgerungen. Folgend ein Beispiel: „We presented the first in-depth study to answer this issue by identifying a process framework including 27 development tasks organized in a three-phase based process framework and then obtaining quantitative and qualitative support through a Webbased survey results. *We provided some recommendations to manage the development process complexity of this class of systems.* The software engineering literature has not observed such a scientific and contextual understanding of the overall development process as it is anticipated to get to know what is new or reiterated“ [28].

Bei den vorenthaltenen Schlussfolgerungen handelt es sich um praktisch anwendbares Wissen. Das Lesen dieses Satzes lässt Lesende nur mit einer genaueren Vorstellung der Schlussfolgerungen zurück, aber nicht mit neuem Wissen. Um dieses zu erhalten, müssen Lesende den gesamten Artikel überfliegen und die passende Stelle dazu finden. Die Schlussfolgerungen sind hier so unkonkret, dass sie nicht verständlich sind. Lesende müssen entweder zusätzlichen Aufwand betreiben oder bleiben ohne Schlussfolgerung zurück. Tritt dies, wie in Abbildung 4, für den einzigen Schlussfolgerungssatz eines Conclusions-Abschnitts auf, oder auch für alle Schlussfolgerungen, erfüllt der Conclusions-Abschnitt nicht seinen Zweck und ist damit ein schlechter Conclusions-Abschnitt.

Signale für Schlussfolgerungen: Eindeutige Hinweise auf eine Schlussfolgerung innerhalb des Satzes sind für eher unerfahrene Lesende von Vorteil. Für erfahrene Lesende sind solche Signale nur zu Beginn des Conclusions-Abschnitts relevant und sonst möglicherweise eher unerwünscht. Für erfahrene Lesende sind Hinweise auf eine Schlussfolgerung hilfreich, wenn die Schlussfolgerung gleich der erste Satz im Abschnitt ist, da der Satz sonst auch als background gelesen werden könnte.

Solche Signale können durch Signalwörter, wie bspw. „we generalize“ oder „our results indicate“, gegeben werden oder die Form von konkrete Empfehlungen haben.

Beispielsweise lässt sich bei folgendem Satz erst durch den nachfolgenden Satz darauf schließen ,ob background oder conclusion-gen besser passt: „*The performance and, ultimately, the applicability of the VGT paradigm strongly depends on the robustness of the underlying image recognition algorithms.* Our results show that state-of-the-art tools have limited portability across devices“ [3]. Ohne Kontext und Fachwissen ist der Satz nicht direkt als Schlussfolgerung identifizierbar. Eindeutiger wäre: „We have found that to raise the performance and , ultimately, the applicability of VGT paradigms the robustness of the underlying image recognition algorithm must be raised.“

Ein positives Beispiel ist folgendes: „Our results can have several implications for researchers and practitioners: [...])“ [31]. Die folgenden Schlussfolgerungen sind jeweils nicht mehr als solche signalisiert, aber durch den zuvor genannten Einleitungssatz und die genutzte Strukturierung klar als solche erkennbar. Das verbessert die Lesbarkeit des Artikels.

Der Faktor Signale beeinflusst die Lesbarkeit, ist aber nicht ausschlaggebend für die Gesamteinschätzung eines Conclusions-Abschnitts.

Nutzung der ersten Person für Schlussfolgerungen: Die Nutzung der ersten Person, im Singular oder Plural, geht meist einher mit eindeutigen Signalen für Schlussfolgerungssätze, ist aber nochmal hervorzuheben, da Autoren sich damit klar hinter die eigenen Schlussfolgerungen stellen. Dies ist vor allem empfehlenswert für Ratschläge, da diese durch die Nutzung der ersten Person kaum verlängert werden. Zur Beschreibung von Gegebenheiten eignet sich die erste Person weniger, da sie immer eine Erweiterung des Satzes ist und diesen damit verlängert und komplexer macht. Dies lässt sich an folgendem Beispiel betrachten: „The phrasing of the messages must take into account their „consumer“ (i.e., the app's user): It is important to avoid technical words (e.g., server, exception, connection error, etc.)“ [24]. Schöner wäre: „We recommend to take the „consumer“ (i.e., the app's user) of each message into account and avoid technical words (i.e. [...]).“ Obwohl dies banal erscheinen mag, wird die erste Person leichter als Schlussfolgerung wahrgenommen und die Autoren übernehmen eindeutig die Verantwortung für ihre Schlussfolgerung. Im Originaltext ist zwar auch erkennbar, dass es sich um eine Schlussfolgerung handelt, ohne Kontext könnte der Satz aber auch bereits bekanntem Wissen entsprechen.

Die Ergebnisse der informellen Studie zur Überprüfung meiner Ergebnisse zeigen einen weiteren Grund auf, neben der Unsicherheit der Autoren, warum Autoren statt der ersten Person die neutrale unbestimmte Schreibweise nutzen: Anscheinend ist dies eine persönliche Vorliebe, die teilweise auf der starken Überzeugung beruht, dass „persönliche Formen gehören nicht in wissenschaftliche Artikel! Wie hört sich das denn an?!“ – Aussage einer teilnehmenden Person der informellen Studie. Das scheint mir nicht für alle Wissenschaftsbereiche zu stimmen. Denn in der Wissenschaft wird auf Genauigkeit viel Wert gelegt und bei qualitativer Forschung, wie meiner, ist ein „ich habe getan“ genauer und näher an der Wahrheit dran als ein „es wurde getan“.

Umgangene Generalisierungen: Sätze oder gar Absätze, die Schlussfolgerungen nicht ausformulieren, sind besonders ärgerlich, weil sie Lesende selbst die Schlussfolgerungen aus den Ergebnissen ziehen lassen. Dies ist eigentlich Aufgabe der Autoren, da diese immer über umfassenderes Wissen über die Studie verfügen.

„We analyzed a total of 276 projects, and found that 85 of them present one or more rotten green tests, accounting for a total of 665 rotten green tests. [...] We see that most of the rotten green tests we detected are caused by context-dependent assertions, although there exist cases of fully rotten tests that hide application bugs“ [2]. Dieser Textauszug sollte nach dem konkreten Kontext des Conclusions-Abschnitts eigentlich die Erkenntnisse der Studie vermitteln. Stattdessen werden nur Ergebnisse präsentiert, Schlussfolgerungen sind nicht vorhanden. Bei eigener Erschließung der Bedeutung des Texts würde ich ihn folgendermaßen formulieren: „Rotten green tests are a relevant problem and can be rougly seen in every third project. To minimize rotten green tests we advise to concentrate on context-dependent assertions.“

Bei diesem Beispiel lässt sich die ausgelassene Generalisierung nur aufgrund des Kontexts aus dem kompletten Conclusions-Abschnitt erschließen, bei dem nach der Beschreibung der Resultate der Schlussfolgerungssatz fehlt. Dies ist für alle fehlende Generalisierungen der Fall. Umgangene Generalisierungen beeinflussen die Gesamtqualität stark und verursachen immer einen schlechten Gesamteindruck, wenn es außerdem keine konkreten Schlussfolgerungen gegeben werden.

Squishy Wording: Beschreibt die Formulierung von Schlussfolgerungen, die eine allgemeine Zustimmung erwirken wollen, indem die Aussagen vage gemacht werden. Dadurch werden Schlussfolgerungen unkonkret. Werden Adjektive genutzt, die einer Erklärung bedürfen, aber im Conclusions-Abschnitt nicht erklärt werden, ist dies weniger schlimm, als wenn der Satz durch Wörter, wie „vielleicht“ oder „can“, unkonkret wird und dadurch sogar seine Bedeutung verändern kann. Adjektive machen Aussagen eher vage. Die meisten Adjektive lassen allerdings eine Tendenz erkennen, auch ohne Definition. Bspw. deutet das Adjektiv „effizienter“ eine positiver Veränderung an. Folgend eine Auflistung von Formulierungen, die Sätze zustimmbarer wirken lassen können, beginnend mit der ungünstigsten Variante.

- Die Verwendung von mutmaßenden Wörtern, die dazu führt, anstatt einer Feststellung nur eine vage Vermutung zu teilen. Damit übermittelt der Satz eine ganz andere Bedeutung als er eigentlich sollte. Wörter wie z.B. „möglicherweise“, „vielleicht“ und „wahrscheinlich“ haben einen solchen Effekt in Schlussfolgerungen.

- Wörter wie „kann“, die Formulierungen erzeugen, die sich wie Existenzbeweise lesen, sind nicht immer als solche gemeint. Dies ist problematisch, da Lesenden hier etwas anderes vermittelt wird, als eigentlich herausgefunden wurde. Solche Sätze zu erkennen ist oft schwer, da nicht jeder Conclusions-Abschnitt genug Informationen enthält, die zeigen, ob es sich um eine Generalisierung oder einen Existenzbeweis handelt.

- Nicht erklärte Adjektive wie „hilfreich“ oder auch „effektiv“ deuten eine Verbesserung an, aber lassen Lesende im Unklaren darüber, wie oder wo genau eine

Verbesserung stattfindet und sind daher ungenau. Da aber davon auszugehen ist, dass diese Adjektive im Hauptteil der Studie erklärt werden, ist dies eine akzeptable Abschwächung von Schlussfolgerungen. Auch weil hier der Sinn des Satzes nicht verändert wird, sondern lediglich Details vorenthalten werden.

„Therefore, these results *may* indicate that OSS communities need to be aware of the diversity of their developers and move towards a more racially and ethnically diverse OSS community“ [60]. Aus diesem Satz geht hervor, dass die Aussage durch die Studienergebnisse begründet ist. Die Autoren scheinen überzeugt genug von ihren Ergebnissen zu sein, dass sie der Meinung sind etwas herausgefunden zu haben, aber sind sich nicht sicher genug, dies auch auszuformuliern. Durch das Wort „may“ wird diese Unsicherheit offensichtlich und die Aussage des Satzes unwichtig, da sie in genau der Form auch vor der Studie getroffen werden konnte. Eindeutiger wäre der Satz ohne das „may“. In diesem Fall verändert ein einzelnes Wort die komplette Aussage eines Satzes. Auch mit „can“ ist dies möglich, bei welchem es schwer wird festzustellen, ob die Schlussfolgerung ein konkreter Existenzbeweis ist oder eine ausgelassene Generalisierung.

Für folgenden Satz ist zwar unklar, was genau „useful and helpful“ bedeutet und welche Auswirkungen dies hat, aber es ist verständlich, dass beide einen positiven Effekt beschreiben und es sich empfiehlt, diese Schlussfolgerung anzuwenden: „We find that documentation with higher content adequacy and conciseness scores is much more *useful and helpful* for code understandability“ [37].

Der Faktor „squishy wording“ beeinflusst die Konkretheit von einzelnen Sätzen sehr stark. Teilnehmende der informellen Studie bevorzugten fast immer den konkreter formulierten Schlussfolgerungssatz. Ein Conclusions-Abschnitt nur mit nur solchen Sätzen der schlechtesten Form wäre nicht als gut zu beurteilen, aber trotzdem nicht einer der schlechtesten Conclusions-Abschnitte. Denn Lesende würden am Ende des Abschnitts wahrscheinlich bemerken, dass alle Schlussfolgerungen sehr vorsichtig formuliert wurden. Wenn alle Schlussfolgerungssätze auffällig vorsichtig formuliert sind, lässt dies die Vermutung zu, dass die Autoren möglicherweise übervorsichtig waren. Lesende könnten die Sätze daher anders interpretieren als sie geschrieben sind und damit entweder korrekt liegen oder falsche Schlüsse ziehen.

Für Forschung bei der es nicht möglich war bestimmtes zu beweisen und für deren Erkenntnisse daher eine vorsichtige Formulierung angebracht wäre, sollte dies verständlich gemacht werden. Autoren sollten dann neben vorsichtigen Formulierungen auch Kontext geben, warum es nicht möglich ist andere Schlussfolgerungen zu ziehen, z.B. in Form von Resultaten oder Einschränkungen.

Komplexer Satzbau: Ein komplexer Satzbau erschwert offensichtlich die Lesbarkeit. Wie sehr ein komplexer Satzbau die Lesbarkeit beeinflusst, ist von der Fähigkeit der Lesenden abhängig. Das bedeutet, dass einfache Sätze die Ergebnisse der Forschung für eine größere Menge an Lesenden zugänglich macht und komplexere Sätze die Lesbarkeit nur für einen Teil von Forschenden verschlechtert. Je komplexer der Satzbau ist, desto kleiner wird die Menge an Lesenden, die den Artikel ohne Problem lesen können.

Ein komplexer Satz kann durch die Menge an Informationen entstehen, die Länge des Satzes, die Anzahl der Verschachtelungen oder auch einfach durch häufige Nutzung desselben neutralen Verweiswortes (z.B.„it“) innerhalb eines Satzes.

„Integrating software supply chain concepts and models to operationalize key variables from social contagion theory to investigate software technology adoption appears to have provided a number of potentially useful insights in the pres-ent case study of two data manipulation technologies within R language“ [56]. ist als ein Satz für mehr Forschende unangenehmer zu lesen als eine Version, die die Informationen des Satzes auf mehrere Sätze verteilt.

Dass die Komplexität von Sätzen die Lesbarkeit beeinflusst und einfache Sätze übermäßig komplexen Sätzen vorzuziehen sind, wurde von allen Teilnehmenden meiner Studie bestätigt. Auch die Quellen [67] und [51] weisen darauf hin, dass Sätze nicht zu kompliziert formuliert werden sollten.

Verworfener Faktor Füllsätze: Dieser Faktor wurde durch meine informelle Studie nicht bestätigt und müsste überarbeitet oder verworfen werden. Dies ist aus Zeitgründen nicht mehr möglich gewesen. Der Studie nach scheint dieser Faktor wenig Einfluss auf die Qualität von Conclusions zu haben und ist stark subjektiv. Konkret wurde innerhalb der Studie klar, dass Füllsätze immer neue Informationen enthalten. Folgend an die Ergebnisse angepasste Hypothesen, die noch zu bewerten sind: Füllsätze enthalten schwer belegbare Aussagen und sind in der Regel direkte Folgen von Schlussfolgerungen, die keine Zukunftsaussichten darstellen. Vermutlich werden Füllsätze dann verwendet, wenn die Schlussfolgerung unpersönlich und wenig überzeugend wirkt. Mit den Füllsätzen versuchen Autoren die Schlussfolgerung relevanter zu machen. Solche Füllsätze erscheinen mir in den meisten Fällen wie langgezogene Schlussfolgerungen, die vermieden werden könnten, wenn konkreter und kompakter forumliert werden würde:

„Our results suggest that incremental build of configurations can reduce build time without trading correctness. *Developers and maintainers can already benefit from this simple approach*“ [66].

Es scheint, dass die Schlussfolgerung in genanntem Beispielsatz so neutral wie möglich und ohne Übernahme der Verantwortung durch die Autoren geschrieben wurden. In solchen Fällen scheint es schwieriger zu sein, den Nutzen und die Bedeutung auszudrücken, die der Autor seinen Schlussfolgerungen beimisst. Eine bessere Formulierung scheint mir hier eine direkte zu sein: "For developers and maintainers we recommend to build configurations incrementally to reduce build time without trading correctness."

5.3.3 Wechselwirkungen zwischen den Faktoren

Zusammenfassend lässt sich sagen, dass der entscheidende Unterschied zwischen guten und schlechten Conclusions-Abschnitten von der Präsentation der Schlussfolgerungen abhängt und damit von der erfolgreichen Wissensvermittlung. Die *Konkretheit von Schlussfolgerungen* über den gesamten Conclusions-Abschnitt hat daher den stärksten Einfluss auf dessen Gesamteindruck. Denn wenn die Schlussfolgerungen unkonkret sind, fehlen Informationen. Unabhängig davon wie schön Sätze dann for-

muliert sind. Wenn die Informationen fehlen, können diese nicht bei den Lesenden ankommen.

Gemäß der eingangs aufgestellten Definition sollten in Conclusions-Abschnitten die wichtigsten Schlussfolgerungen zu finden sein, da diese die Erkenntnisse der Studie und damit ihr Wissen bündeln. Alles andere, was im letzten Abschnitt steht, sind nach dieser Annahme unverbindliche Zusatzinformation, die relevant sein können, aber niemals wichtiger. Daher sind Abschnitte, die keine Schlussfolgerungen enthalten, immer schlechter, als solche, die dies tun. Das konkrete ausformulieren von Schlussfolgerungen kann umgangen werden, indem nur die Ergebnisse genannt werden und auf eine Generalisierung verzichtet wird. Das kann erreicht werden, indem auf mögliche Schlussfolgerungen verwiesen wird ohne diese genau zu verorten oder indem Schlussfolgerungen unkonkret formuliert werden.

Darüber hinaus kann als allgemeine Richtlinie gesagt werden, dass Conclusions-Abschnitte nur dann besonders gut sind, wenn sie leicht lesbar sind und Lesende ohne viel zusätzlichen Aufwand die Schlussfolgerungen direkt vermittelt bekommen. Konkretheit beeinflusst die Qualität des Conclusions-Abschnitts stärker als Lesbarkeit. Die Lesbarkeit eines Conclusions-Abschnitts beeinflusst die Qualität nur selten stark genug, um aus einem Abschnitt mit guter Konkretheit einen insgesamt schlechten Abschnitt zu machen. Dafür müsste die Lesbarkeit des Artikels so schlecht sein, dass die Schlussfolgerungen fast unverständich sind. So ein Beispiel habe ich während meiner Analyse nicht gesehen. Keiner der analysierten Artikel fällt in diese Kategorie. Durch vorherige Prüfung von wissenschaftlichen Artikeln ist es außerdem unwahrscheinlich. dass dieser Fall eintritt. Speziell die Faktoren *Strukturiertheit* und *Signale für Schlussfolgerungen* können einen Gesamteindruck nur leicht verändern. Teilnehmende der Studie legten auf Signale für Schlussfolgerungen je nach persönlicher Vorliebe mehr oder weniger Wert. Die Strukturiertheit eines Artikels, kann diesen verbessern, aber nicht retten. Ein sonst sehr unkonkreter Artikel, der sehr gut strukturiert ist, ist insgesamt als schlechter Artikel einzuschätzen, da die Schlussfolgerungen zu unkonkret sind. Bei *komplexen Satzbau* waren sich alle Teilnehmenden meiner informellen Studie einig, dass dieser als eher zu vermeiden angesehen werden sollte. Die Bewertungen der Teilnehmenden deutet darauf hin, dass die Komplexität des Satzes mehr Einfluss auf die Lesbarkeit hat als Signalwörter für Schlussfolgerungen. Ein extrem komplex geschriebener Conclusions-Abschnitt wäre nicht für alle Forschenden unverständlich, für einen kleinen Rest an Forschenden bliebe die Lesbarkeit unverändert. Dies verkleinert allerdings die Menge an Lesenden, die Nutzen aus dem Conclusions-Abschnitt ziehen können, was nachteilig ist.

Der Faktor *Anteile Schlussfolgerung vs. Einleitung* beeinflusst neben der Lesbarkeit auch leicht die Konkretheit von Schlussfolgerungen. Wenn zu wenig Schlussfolgerungsinformationen im Gegensatz zu einleitenden Informationen im Abschnitt vorhanden sind, z.B. ein Drittel gegen zwei Drittel, erscheint der Abschnitt nicht mehr seinen Zweck zu erfüllen. Je mehr für die Schlussfolgerung irrelevanten Informationen genannt werden und je weniger Schlussfolgerungen selber genannt werden, desto weniger scheint die Forschungsarbeit des Artikels zielführend gewesen zu sein.

5.3.4 Einschränkungen der Erkenntnisse

Dieser Abschnitt beschreibt zu beachtende Einschränkungen bzgl. meiner gefundenen Qualitätsfaktoren und ggf. Maßnahmen, die ich ergriffen habe, um diese Einschränkungen zu beheben.

Die gefundenen Faktoren sind sehr wahrscheinlich von meiner eigenen Voreinstellung beeinflusst, da sie auf subjektiven Einschätzungen beruhen. Daher versuche ich folgend zu benennen, was einen Einfluss auf meine Beurteilung gehabt haben könnte.

- Im Lesen wissenschaftlicher Artikel bin ich nicht besonders geübt. Im Masterstudium habe ich wahrscheinlich ca. drei bis zehn Artikel pro Semester vollständig gelesen, mindestens zwanzig Artikel punktuell. Dabei bin ich meist der Reihenfolge gefolgt, mit dem Abstract zu Beginnen, dann den Conclusions-Abschnitt zu lesen und zuletzt den Rest des Artikels. Um Artikel gründlich und vollständig zu lesen, habe ich meine Notizen nutzen müssen, habe also noch keine Systematik zum Verstehen verinnerlicht.

- Meine Fähigkeit Schlussfolgerungen zu erkennen war zu Beginn der Arbeit nicht ausgeprägt. Beispielsweise hatte ich die Eindeutigkeit von Schlussfolgerungssätzen bzgl. ihrer Funktion als Schlussfolgerung als eher wichtiges Qualitätskritrium angesehen, da solche Sätze mit eindeutigen Signalen, dass eine Schlussfolgerung präsentiert wird, einfacher für mich zu verstehen sind. Dass ein solcher Satz dadurch länger wäre als ohne diese Signale, ist für mich eher unwichtig. Aus den Ergebnissen meiner informellen Studie ist allerdings hervorgegangen, dass andere Forscher diese Vorliebe nicht teilen und eine Signalisierung von Schlussfolgerungen nicht immer erwünscht ist.

- Meine Fähigkeit Englisch zu lesen und zu verstehen ist im Laufe der Arbeit gestiegen. Lange und komplexe Sätze benötigten mehr Aufwand. Daher erscheinen mir Artikel mit solchen Sätzen tendenziell eher als schlecht.

- Da ich in den meisten der Themen der von mir analysierten Artikel kein Expertenwissen besitze, habe ich inhaltlich vieles nur oberflächlich verstehen können. Mir ist aufgefallen, dass mir das Codieren von Conclusions-Abschnitten über mir bekanntere Themenbereiche einfacher von der Hand ging. Damit waren Artikel mir bekannter Themen auch einfacher zu verstehen und erschienen daher wahrscheinlich auch tendenziell als besser.

GTM-Theorien kommen aus den Daten und können daher auch nur so gut sein, wie die zugrundeliegenden Daten [36]. Die analysierten Conclusions-Abschnitte stammten alle aus empirischen Artikeln desselben Jahres aus dem SWE und sind zufällig ausgewählte Artikel. Damit ist die Datengrundlage gut genug um Qualitätsmerkmale von häufig auftretenden Conclusions-Abschnitten zu erforschen. Eher selten auftretende gute Conclusions-Abschnitte können allerdings in der Datenbasis fehlen.

Folgende Einschränkungen bzgl. des Codierens sind zu beachten:

- Durch menschliche Fehler, kann vergessen werden Codes zu vergeben. Wenn ich z.B. den strukturellen Inhalt codierte, übersehe ich manchmal Abkürzungen

oder Satzteile, die codiert werden müssen. Ich denke, dass das an den unterschiedlichen Granularitäten lag, mit denen ich codierte, und zum Teil an den unterschiedlichen Perspektiven, die codiert wurden. Zum Beispiel codiere ich nicht nur, was der Inhalt ist, sondern auch noch, ob es erklärte oder unerklärte Abkürzungen im Satz gibt.

- Da ich größtenteils alleine codiert habe, ist die Möglichkeit höher, dass Codes fehlerhaft vergeben wurden. Die Anzahl solcher falsch vergebenen Codes sollte aber klein sein, da ich nach einem Monat Pause alle meine Artikel erneut codierte und damit praktisch alle Artikel doppelt codiert habe.
- Es gibt Sätze, die ohne Informationen aus dem Hauptteil des Artikels schwer zu codieren sind. Insgesamt habe ich so einen Code in 20 Artikeln für jeweils ein bis drei Sätze vergeben. Das bedeutet, jeder zweite Artikel hat einen solchen schwer zu codierenden Satz.
- Da die gefundenen Faktoren nicht auf der genauen Vergabe von Codes beruhen, ist dies für diese kaum relevant. Für die Ergebnisse der quantitativen Analyse der Studierenden ist dies eher relevanter.
- Obwohl dies dort die Korrektheit der Codes beeinflusst, lässt dies eher Teilergebnisse zum Aufbau von Conclusions-Abschnitten ungenauer werden. Dies verändert nicht die Aussage, dass rein strukturell inhaltliche Codes und objektive Codes nicht genügen um die Qualität von Conclusions-Abschnitten zu beurteilen.

Die beiden ersten Codierungsprobleme treten in einer Stichprobe von sechs Conclusions-Abschnitten nicht bzw. selten auf. Es wurden keine Codierungen falsch vergeben Allerdings wurde in einem Abschnitt eine Codierung vergessen: Es fehlt die Codierung einer Abkürzung. Außerdem wurde an zwei Stellen in drei Abschnitten vergessen der Code zu vergeben, welcher Sätze markiert, die ohne Informationen aus dem Hauptteil des Artikels, schwer zu codieren sind.

6 Fazit

Im Folgenden werde ich zuerst meine Erkenntnisse bzgl. des Forschungsthemas zusammenfassen, dann auf mein persönliches Vorgehen eingehen und zuletzt einen Ausblick geben.

Ziel dieser Masterarbeit war es, die Qualität von wissenschaftlichen SWE-Artikeln zu untersuchen, genauer die Qualität der Conclusions-Abschnitte zu beurteilen. Annahme dabei war, dass Schlussfolgerungen ein essentieller Teil von Conclusions-Abschnitten sind. Diese Annahme und die bottom-up Herangehensweise, schränkte meine Forschung so ein, dass ich keine allgemein gültigen Qualitäts Kriterien für Conclusions-Abschnitte finden konnte. Selbst detailliertere Codes zur Qualität der analysierten Abschnitte, wären nicht ausreichend gewesen. Denn es ist nicht garantiert, dass in meiner Datenmenge eher seltener vorkommende Ausprägungen von Conclusions-Abschnitten vorkommen. Außerdem scheinen neben Schlussfolgerungen auch andere Informationen für Conclusions-Abschnitte essentiell [33]. Demnach sind z.B. fhope

Sätze keine ungenau formulierten Ausblicke, sondern Beschreibungen der Vision der Autoren des Artikels, die helfen, die Forschung in einen größeren Kontext einzuordnen.

Einzelnen Hinweisen bezüglich des Zwecks und der Qualität von Conclusions-Abschnitten konnte ich erarbeiten:

- Schlussfolgerungen sind der wichtigste Teil einer Forschung und damit auch von wissenschaftlichen Publikationen. Diese Aussage beschreibt die Wissenschaftstheorie und auch im SWE wird dies vertreten. Folglich lässt sich argumentieren, dass der letzte Abschnitt, der Conclusions-Abschnitt, die wichtigsten der Schlussfolgerungen mitteilen sollte. Diese Aussage konnte ich im SWE nicht bestätigen. Obwohl im SWE wenig konkret über einen Conclusions-Abschnitt gesprochen wird, ist dieser in jedem aktuellen Artikel zu finden. Dies deutet darauf hin, dass der Abschnitt entweder durch die Konferenzen vorgegeben wird oder fest in der Vorstellung von Forschenden verankert ist, ohne eine genaue Definition.

- Um die gesamte Qualität eines Artikels zu beurteilen, ist es sinnvoll, die Qualität von Schlussfolgerungen innerhalb des ganzen Artikels zu betrachten. Dafür müsste neben dem Conclusions-Abschnitt auch immer der Diskussion-Abschnitt (wenn auch anders benannt) betrachtet werden.

- Es gibt zu wenige SWE-Artikel, die sich mit dem Schreiben und Veröffentlichen wissenschaftlicher Arbeiten befassen. Man könnte daher annehmen, dass wissenschaftliche Arbeiten im SWE viele verschiedene Formen annehmen können, und zwar jeweils die Form, die dem Autor am geeignetsten erscheint. Dies ist jedoch nicht der Fall, da wissenschaftliche Arbeiten stark durch die Erwartungen der wissenschaftlichen Gemeinschaft vorgegeben und standardisiert sind. Ein weiterer Teil wird von den Konferenzen vorgegeben, die die Artikel publizieren. Dies führt zu einem relativ einheitlichen Erscheinungsbild aller Artikel. Interessanterweise scheint dies auch dazu zu führen, dass jede Arbeit einen Conclusions-Abschnitt hat, obwohl ein solcher Abschnitt wenig diskutiert wird. Ich vermute, dass dies zu einem achtlosen Schreiben führt. Bestätigt wird dies dadurch, dass innerhalb meiner analysierten Artikel ca. jeder Dritte Artikel, trotz eigener Ankündigung in der Gliederungen, in Conclusions-Abschnitten keine Schlussfolgerungen enthalten.

- Es ist nur schwer möglich die Intentionen bzw. mentalen Einstellungen von Autoren allein anhand von Conclusions-Abschnitten zu beurteilen. Veröffentlichte wissenschaftliche Artikel sind weit entfernt von der eigentlichen Ausdrucksweise der Autoren. Solche Artikel werden nie umgangssprachlich formuliert und je nach Review Prozess der jeweiligen Konferenz mehrfach überarbeitet und standardisiert. Die Form die der veröffentlichte Artikel hat, entspricht damit selten einer natürliche Ausdrucksweise aus der mentale Einstellungen gelesen werden können. Auch zu beachten ist, dass bei einem Autoren Team, pro Artikel, Autoren verschiedene Einstellungen haben können, die sich nur anhand des Artikels nicht auf einzelne Personen zurückführen lassen.

- Allein anhand von strukturellen Inhalt-Codes ist die Qualität von Conclusions-Abschnitten nicht zu beurteilen. Dies hat die Erpobung des Codebuchs veranschaulicht. Diese Codes können genutzt werden, um sich innerhalb von Conclusions-Abschnitten zurecht zu finden und irrelevante Stellen zu identifizieren und Tendenzen in der Qualität zu erkennen. Conclusions-Abschnitte mit einem größeren Textanteil an Schlussfolgerungen als Hintergrundinformationen sind gute Conclusions-Abschnitte, wenn die einzelnen Schlussfolgerungssätze konkret genug sind.

Insgesamt habe ich keine detaillierte Antwort darauf finden können, wie die Qualität von Conclusions-Abschnitten beurteilt werden kann. Daher ist es auch nicht möglich, nach dieser Arbeit eine ähnliche quantifizierende Studie zum aktuellen Stand der Qualität von Conclusions-Abschnitten im SWE zu machen, wie die qabstracts-Studie es für Abstracts umsetzt. Ich denke, dass eine quantifizierende Studie auf Basis eines mit der gewählten Methode erstellten Codebuchs nicht ausreichend ist, um eine bessere Qualität von Conclusions-Abschnitten im SWE zu erreichen. Selbst wenn die Methode geeignet wäre, Qualitätsfaktoren umfassend zu identifizieren, würden die Qualitäts-Faktoren und die daraus abgeleiteten Empfehlungen zum Schreiben immer nur so gut sein wie die Daten, aus denen sie stammen, also aus konkreten publizierten Conclusions-Abschnitten. Bei diesen scheint es deutlich mehr schlechte als gute Abschnitte zu geben.

Schöner wären Richtlinien zu Conclusions-Abschnitten, die mit Hilfe der Wissenschaftstheorie und -rhetorik erarbeitet und auch begründet werden könnten. Der dabei entstehende Vorschlag zu Conclusions-Abschnitten würde von einem begründeten Ideal ausgehen und nicht von einer Menge an mehr oder weniger aufmerksam geschriebenen Abschnitten, die an subjektive Ideale, Vorgaben durch Konferenzen und erlernte Vorgaben durch das viele Lesen anderer Artikel stammen.

6.1 Selbstreflexion

Insgesamt erschien mir mein erster Teil der Arbeit eher unsystematisch. Da ich mich in die Techniken der GTM und das Thema einarbeiten musste, fiel es mir schwer, mich zu organisieren und damit meine gesamten Fähigkeiten auszuschöpfen. Mir ist viel Zeit verloren gegangen durch den Fokus darauf, eine genaue Definition für Schlussfolgerungen zu finden und genau zu begreifen. Um dieses Ziel zu erreichen, versuchte ich, meine theoretische Sensitivität hauptsächlich durch das Codieren weiterer Quellen zu erhöhen. Im zweiten Teil meiner Forschung fokussierte ich mich, aufgrund der bereits vorhandenen einfacheren Codes und der Sicherheit bzgl. Aussagen objektiver Codes, darauf, Qualität mit Hilfe von objektiven Codes zu bestimmen. Dies stellte sich als nur sehr begrenzt möglich heraus. Erst im dritten Teil meiner Forschung kam ich auf subjektive Einschätzungen zurück und wandte gezielt Techniken der GTM an, um meine theoretische Sensitivität zu erhöhen. Aus diesem Teil der Arbeit stammen meine vorgestellten Qualitätsfaktoren.

Im folgenden habe ich eine Liste an persönlicher Schwierigkeiten, die meine Forschungsarbeit erschwert haben. Im Anschluss findet sich eine Liste mit Hinweisen, die mir zu Beginn meiner Forschung gefehlt haben und die für andere Forschende möglicherweise hilfreich sein könnte.

Über mich persönlich habe ich im Laufe der Masterarbeit folgendes gelernt:

- Zwischenziele, Ordnung und Systematik sind wichtig, damit ich mich bei meiner Arbeit wohlfühle und motiviert bin. Dies habe ich zu Beginn meiner Forschung schlecht umsetzen können, da ich meine theoretische Sensitivität nicht durch systematisches Anwenden verschiedener Techniken versuchte zu erhöhen, sondern nur durch das Betrachten neuer Quellen in wiederholender Art und Weise. Wahrscheinlich hätte mir ein periodisches Zurücktreten und Zusammenfassen meiner bisherigen Ergebnisse helfen können, schneller zu realisieren, dass inhaltliche Codes nicht zu meinem Ziel führen.

- Ich arbeite besser im Team. Es fällt mir leicht, Zwischenziele zu setzen und kritisch zu denken, wenn ich regelmäßig mit jemandem besprechen kann. Ohne diesen Anstoß fällt es mir schwerer meine Gedanken und Fortschritte regelmäßig auszuformulieren und zu sortieren. Ich bin leichter zu motivieren durch externe Motivationsanstöße, wie Absprachen, Wettbewerbe, Wertschätzung der Arbeit und einen Gemeinschaftsgedanken. Solche Motivation war aus verschiedenen Gründen kaum möglich während dieser Masterarbeit. Ich hatte nicht genügend Möglichkeiten, mit Leuten regelmäßig über meine Forschungsarbeit zu sprechen. Mir fehlte eine Person regelmäßig und häufig zuhören und Feedback geben konnte, ohne dass immer alles von vorne erklärt werden musste. So wurde ich nicht daran erinnert regelmäßig meine Gedanken zu ordnen, um über meine Fortschritte zu berichten. Nachdem ich dies bemerkt hatte, habe ich versucht Abhilfe zu schaffen, indem ich immer selbst E-Mails an meinen Betreuer geschrieben habe, in denen ich formuliert habe, was ich gemacht habe und was ich vorhabe, ohne diese abzuschicken. Da es mir später leichter fiel, ein Tagebuch zu führen, ging ich dazu über, jeden Tag mit einem Tagebucheintrag bzgl. meiner Forschung zu beginnen.

  Die Arbeit mit den Studierenden und die gemeinsame Organisation des Projekts war eine sehr willkommene Abwechslung, in der mir die Arbeit leichter fiel. Ich war motivierter und hatte mehr Ideen.

- Ich bin ein Mensch, der sich nach dem Sprichwort „erst die Arbeit, dann das Vergnügen“ organisiert. Das war bei meiner Forschung hinderlich, weil ich neue gut erscheinende Ideen nicht direkt verfolgte sondern erst liegen ließ. Dadurch verlor ich einige Ansätze zu Hypothesen. Nach dieser Realisation habe ich versucht, mehr Impulsen für Hypothesen direkt nachzugehen und hatte das Gefühl, mehr über die Qualität von Schlussfolgerungen zu lernen als zuvor. Außerdem war ich wesentlich motivierter. Dazu ist allerdings anzumerken, dass ich mein Verhalten erst zum Ende der Analysephase bemerkt und geändert habe. Das bedeutet, dass ich schon viel mehr Theoretische Sensitivität über Conclusions-Abschnitte angesammelt hatte als zuvor. Dies hat das Finden von Hinweisen zur Qualität zu Conclusions-Abschnitten sicherlich auch begünstigt.

- Ich habe eher wenig Selbstvertrauen in meine eigene Arbeit. Dies führt i.d.R. dazu, dass ich Arbeiten genauer mache und versuche, möglichst perfekt zu beenden. Bei der Masterarbeit hat dies allerdings konkret dazu geführt, dass es für

mich viel schwieriger war, meine eigenen Hypothesen zu verfolgen und mich auf mein neu erlangtes Wissen zu verlassen. Aus diesem Grund habe ich mich wahrscheinlich extrem lange mit dem Verständnis von Schlussfolgerungssätzen beschäftigt, bevor ich mich darauf verlassen habe. In Kombination mit dem alleine Arbeiten hat dies dazu geführt, dass es mir schwer fiel, zwischen guten und schlechten Ideen zu unterscheiden.

Folgend die Liste an Empfehlungen für einsteigende qualitive Forschende:

- **An die ersten Codes beim offenen Codieren wenig Ansprüche haben:** Die initiale Codierungsphase ist eine, die stark auf Kreativität und Intuition der Forschenden beruht. Die dabei entstehenden Codes werden im Laufe der Forschung weiterverarbeitet. Das bedeutet, diese Codes sieht i.d.R. niemand anderes als die Forschenden selbst und die Codes können wichtige Ansätze für die Forschung sein.

- **Systematisch vorgehen:** GTM ist keine „Durststrecke mit unbekanntem Ende“, sondern ein langer unbekannter Weg zu einem Ziel. Auch wenn GTM ein kreativer Prozess ist, sollte man bekannte Methoden anwenden und sich in regelmäßigen Abständen die Zeit nehmen, einen Schritt zurückzutreten und zu konsolidieren, auch wenn man das Gefühl hat, keine neuen Erkenntnisse gewonnen zu haben. Man sollte sich Zwischenziele und Meilensteine setzen, auch wenn sie noch so klein erscheinen. Z.B das periodische Zurücktreten und Zusammenfassen als Zwischenschritt und das erneute Anwenden von Micro-Coding-Methoden, eventuell mit einer anschließenden Evaluation, warum die Ergebnisse einem nicht weiterhelfen und wie man das ändern könnte.

- **Codierungssitzungen, nicht nur alleine:** Eine weitere Empfehlung für Personen, die unerfahren im Codieren sind, wäre, eine Codierungssitzung mit einer anderen Person gemeinsam durchzuführen. Zusätzlich dazu wäre eine angeleitete Sitzung hilfreich, in der die systematischen Analysemethoden von Strauss und Corbin in der Anwendung gezeigt werden können. Im Laufe der Arbeit habe ich die Möglichkeiten, mit anderen zu codieren oder Codes zu vergleichen, sehr geschätzt. Nach solchen Sitzungen hatte ich immer neue Ideen oder zumindest ein besseres Verständnis der Codes.

- **Betrachtungsebenen bewusst anwenden:** Der Gedanke der verschiedenen Betrachtungsebenen sollte im Kopf behalten werden, wenn analysiert wird. Ich habe erst viel zu spät realisiert, dass ich unterschiedliche Perspektiven eines Textes gezielt und systematisch analysieren sollte.

6.2 Ausblick

Mit den jetzigen Erkenntnissen der Studie ist es nicht möglich, eine quantifizierende Studie auf Basis des Codebuchs zu machen, die einen Überblick über die Qualität von Conclusions-Abschnitten gibt. Dazu fehlt noch ein ausgearbeitetes Codebuch, das Qualitätsmängel quantifizierbar macht.

Sobald es ein solches Codebuch geben würde, mit dem die Qualität von Conclusions-Abschnitten ermittelt werden kann, wäre es möglich daraus eine Künstliche

Intelligenz (KI) zu entwickeln, mit Hilfe derer Conclusions-Abschnitte automatisch codiert werden könnten. Damit könnten Conclusions-Abschnitte schnell automatisch beurteilt werden und z.B. als Hilfsmittel genutzt werden, bevor Artikel publiziert werden. Dies ließe sich auch mit maschinellem Lernen vom SWE auf andere Wissenschaftsbereiche ausweiten.

Mit Hilfe eines solchen Werkzeugs könnten leichter große Mengen an Conclusions-Abschnitten analysiert werden und es könnte als eine Art Überprüfung beim Schreiben solcher Abschnitte genutzt werden.

So ein Werkzeug könnte aus zwei Teilen bestehen. Ein Teil wäre eine KI, die codiert, der andere Teil ein Algorithmus der aus codierten Artikeln auf die Gesamtqualität schließt. Solch eine codierende KI zu bauen, ist theoretisch möglich und würde grob einer Multi-Class-Klassifizierung von Texten entsprechen. Ein Beispiel für eine absatzweise ähnliche KI ist in der Dokumentation bzw. dem Tutorial-Teil des KI-Frameworks Tensorflow beschrieben [27]. Zum Codieren aller Sätze innerhalb eines Conclusions-Abschnitts wäre sicherlich eine komplexere KI nötig, die für jeden Satz auch den Kontext des Textes mit einbezieht. Ein möglicher Ansatz ist ein KI Algorithmus, welcher Sätze innerhalb eines Dokuments vorgegeben Kategorien zuordnet [75]. Dieser KI-Algorithmus nutzt sowohl zuvor per Hand codierte Daten, als auch nicht codierte Daten. Dies würde die Datenmenge die zuvor codiert werden muss verkleinern. Trotzdem bestehen Probleme bei einer solchen Umsetzung: Zum einen ist die Anzahl der Codes, die voraussichtlich benötigt werden um die Qualität zu beurteilen wahrscheinlich relativ groß. Das ist für das Trainieren einer KI nachteilig, da dies mehr benötigte codierte Daten, zum Lernen, und mehr benötigte Trainingskapazitäten bedeutet – möglicherweise zu viel. Als Vergleich: In dem Artikel wurden 20 Klassen genutzt[75], das Codebuch hält momentan allein schon 30 Inhalts-Codes. Zum anderen sind die in dem Artikel verwendeten Klassen einfacher voneinander zu unterscheiden als die Codes des Codebuchs. Denn in dem Artikel wird zwischen Themengebieten unterschieden. Diese können schon allein anhand des verwendeten Vokabulars unterschieden werden. Die zu unterscheidenen Codes jeden Abschnitts besprechen immer dasselbe Thema und sind teilweise nur aus dem Kontext heraus differenzierbar.

Um die Qualität von Conclusions-Abschnitten besser zu verstehen und einen Anreiz für Forschende zu schaffen gute Conclusions-Abschnitte zu schreiben, könnte es eine solide Grundlage sein, eine Analyse innerhalb der Wissenschaftstheorie durchzuführen. Diese sollte herausfinden, was genau der Zweck wissenschaftlicher empirischer Arbeiten ist. Basierend darauf könnte gemeinsam mit Wissenschaftrhetorikern, Wissenschaftltern der Wissenschaftskommunikation, Psychologen und Didaktikern ein Konzept erarbeitet werden, wie dieser Zweck am besten in Artikeln erreicht werden kann. Die Beteiligung von mindestens einem der genannten Fachbereiche erscheint mir sinnvoll, da wissenschaftliche Artikel eine Art der Kommunikation von Forschenden sind, und alle drei der Fachbereiche sich mit Kommunikation zwischen Menschen beschäftigen. Die dabei entstehenden Empfehlungen zum Schreiben von Artikeln und speziell Conclusions-Abschnitten, sofern diese als relevant identifiziert werden, könnten genutzt werden, um Qualitätskriterien für Schlussfolgerungen zu finden. Mit diesen Qualitätskriterien könnte eine quantifizierende Studie durchge-

führt werden, die den Stand der Qualität zeigt, der mit größter Wahrscheinlichkeit nicht gut wäre, so dass den Hinweisen zum Schreiben guter Conclusions-Abschnitte mehr Aufmerksamkeit zukommen würde.

Wahrscheinlich weniger aufwendig, um Qualitätskriterien zu finden, wäre es, sich beim qualitativen Analysieren, auf eine rhetorische Perspektive zu fokussieren, da ich diese noch nicht ausgeschöpft habe.

Zusätzlich könnten Conclusions-Abschnitten aus den Sozial- und Wirtschaftswissenschaften betrachtet werden, deren Themenfelder teilweise ähnlicher sind zu empirischer Forschung in der Informatik als es die Ingenieurwissenschaften sind zu denen das SWE anfänglich gezählt wurde.

Literatur

[1] William P. Alston. „The Place of the Explanation of Particular Facts in Science“. In: *Philosophy of Science* 38.1 (1971), S. 13–34. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/186759 (besucht am 04. 09. 2023).

[2] Vincent Aranega u. a. „Rotten Green tests in Java, Pharo and python“. In: *Empirical Software Engineering* 26.6 (2021). DOI: 10.1007/s10664-021-10016-2.

[3] Luca Ardito u. a. „Feature matching-based approaches to improve the robustness of Android Visual GUI testing“. In: *ACM Transactions on Software Engineering and Methodology* 31.2 (2021), S. 1–32. DOI: 10.1145/3477427.

[4] Sebastian Baltes und Paul Ralph. „Sampling in software engineering research: A critical review and guidelines“. In: *Empirical Software Engineering* 27.4 (2022). DOI: 10.1007/s10664-021-10072-8.

[5] Ingolf Bamberger und Ludwig Mair. „Die Delphi-Methode in der Praxis: Ergebnisse einer exploratorischen Untersuchung zu Einsatzbereichen und Anwendererfahrungen“. In: *Management International Review* 16.2 (1976), S. 81–91. ISSN: 0025181X. URL: http://www.jstor.org/stable/40227267 (besucht am 27. 10. 2023).

[6] Ann Barcomb u. a. „Managing episodic volunteers in free/libre/open source software communities“. In: *IEEE Transactions on Software Engineering* 48.1 (2022), S. 260–277. DOI: 10.1109/tse.2020.2985093.

[7] Marius-Tudor Benea. „Tips for Writing a Good Scientific Paper“. In: *XRDS* 21.4 (Juli 2015), S. 11. ISSN: 1528-4972. DOI: 10.1145/2788532. URL: https://doi.org/10.1145/2788532.

[8] A. Cornelius Benjamin. „The Mystery of Scientific Discovery“. In: *Philosophy of Science* 1.2 (1934), S. 224–236. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/184391 (besucht am 05. 09. 2023).

[9] Henry Evelyn Bliss. „The System of the Sciences and the Organization of Knowledge“. In: *Philosophy of Science* 2.1 (1935), S. 86–103. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/184558 (besucht am 04. 09. 2023).

[10] *Britannica Dictionary definition of conclusion*. 2023. URL: https://www.britannica.com/dictionary/conclusion.

[11] Steven R Brown. „A Primer on Q Methodology“. In: *Operant Subjectivity* 16 (Jan. 1993). DOI: 10.22488/okstate.93.100504.

[12] Martin Carrier. *nbsp;Interview: Wissenschaftsphilosophie: Von der Wissenschaftstheorie zur Wissenschaftsphilosophie*. März 2008. URL: https://www.information-philosophie.de/?a=1&t=939&n=2&y=4&c=68.

[13] Jeffrey Carver. „The use of grounded theory in empirical software engineering“. In: *Empirical Software Engineering Issues. Critical Assessment and Future Directions: International Workshop, Dagstuhl Castle, Germany, June 26-30, 2006. Revised Papers*. Springer. 2007, S. 42–42.

[14] Tim Clark und June Kinoshita. „Alzforum and SWAN: The Present and Future of Scientific Web Communities“. In: *Briefings in bioinformatics* 8 (Juni 2007), S. 163–71. DOI: 10.1093/bib/bbm012.

[15] J. Corbin und A. Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, 2014. ISBN: 9781483315683. URL: https://books.google.de/books?id=hZ6kBQAAQBAJ.

[16] Magnus Dahlgren. *10 check-in exercises for agile retrospectives - Magnus Dahlgren*. Sep. 2019. URL: https://magnusd.cc/check-in-exercises-agile-retrospectives.html (besucht am 28. 10. 2023).

[17] Robert A. Day. „How to write a scientific paper“. In: *IEEE Transactions on Professional Communication* PC-20.1 (1977), S. 32–37. DOI: 10.1109/TPC.1977.6594174.

[18] Robert A. Day und barbara Gastel. *How to write and publish a scientific paper, 8th Edition*. Greenwood, 2016.

[19] Kris Deering und Jo Williams. „Approaches to reviewing the literature in grounded theory: a framework“. In: *Nurse Researcher* 28.4 (Dez. 2020), S. 9–15. DOI: 10.7748/nr.2020.e1752. URL: https://doi.org/10.7748/nr.2020.e1752.

[20] Wolfgang Deppert. *Theorie der Wissenschaft Band 1: Die Systematik der wissenschaft*. Springer Fachmedien Wiesbaden GmbH, 2016.

[21] Chris Drew. *How to read a journal article in 7 steps*. Jan. 2023. URL: https://helpfulprofessor.com/journal-articles/.

[22] Charles G Durbin. „How to read a scientific research paper“. In: *Respiratory care* 54.10 (2009), S. 1366–1371.

[23] *Erkenntnis*. URL: https://www.springer.com/journal/10670 (besucht am 29. 10. 2023).

[24] Camilo Escobar-Velásquez u. a. „Studying eventual connectivity issues in Android apps“. In: *Empirical Software Engineering* 27.1 (2021). DOI: 10.1007/s10664-021-10020-6.

[25] Herman A. Estrin. „How to write for scientific and technical journals“. In: *IEEE Transactions on Professional Communication* PC-25.1 (1982), S. 32–33. DOI: 10.1109/TPC.1982.6447742.

[26] *European Journal for Philosophy of Science*. URL: https://www.springer.com/journal/13194 (besucht am 29. 10. 2023).

[27] *Exercise: multi-class classification on Stack Overflow questions*. Sep. 2023. URL: https://www.tensorflow.org/tutorials/keras/text_classification#exercise_multi-class_classification_on_stack_overflow_questions (besucht am 29. 10. 2023).

[28] Mahdi Fahmideh u. a. „Software engineering for internet of things: The practitioners' perspective". In: *IEEE Transactions on Software Engineering* 48.8 (2022), S. 2857–2878. DOI: 10.1109/tse.2021.3070692.

[29] João Fernandes. „Authorship trends in software engineering". In: *Scientometrics* 101 (Okt. 2014), S. 257–271. DOI: 10.1007/s11192-014-1331-6.

[30] Juan Manuel Florez u. a. „An empirical study of data constraint implementations in Java". In: *Empirical Software Engineering* 27.5 (2022). DOI: 10.1007/s10664-022-10175-w.

[31] Stefanos Georgiou u. a. „Green AI". In: *Proceedings of the 44th International Conference on Software Engineering* (2022). DOI: 10.1145/3510003.3510221.

[32] Barney G. Glaser. *Glaser. basics of grounded theory analysis*. Sociology Press, 1992.

[33] Hilary Glasman-Deal. *Science research writing: For native and non-native speakers of English*. World Scientific, 2021.

[34] Lina Gong u. a. „Revisiting the impact of dependency network metrics on software defect prediction". In: *IEEE Transactions on Software Engineering* (2021). DOI: 10.1109/tse.2021.3131950.

[35] Alan G. Gross, Joseph E. Harmon und Michael S. Reidy. *Communicating science: The scientific article from the 17th century to the present*. Oxford University Press, 2002. Kap. 8–9.

[36] Rashina Hoda, James Noble und Stuart Marshall. „Developing a grounded theory to explain the practices of self-organizing Agile teams". In: *Empirical Software Engineering* 17 (2012), S. 609–639.

[37] Xing Hu u. a. „Correlating automated and human evaluation of code documentation generation quality". In: *ACM Transactions on Software Engineering and Methodology* 31.4 (2022), S. 1–28. DOI: 10.1145/3502853.

[38] W. Hume-Rothery. „A note on the writing of scientific papers". In: *IRE Transactions on Engineering Writing and Speech* 3.2 (1960), S. 43–44. DOI: 10.1109/TEWS.1960.6594509.

[39] *International Studies in the Philosophy of Science*. URL: https://www.tandfonline.com/toc/cisp20/current (besucht am 29. 10. 2023).

[40] Andreas Jedlitschka und Dietmar Pfahl. „Reporting guidelines for controlled experiments in software engineering". In: *2005 International Symposium on Empirical Software Engineering*. IEEE. 2005, 10–pp.

[41] *Journal for General Philosophy of Science*. URL: https://www.springer.com/journal/10838/ (besucht am 29. 10. 2023).

[42] Kim Sung Eun Kang Gwansuk. „How to write an original article in medicine and medical science". In: *kmj* 37.2 (2022), S. 96–101. DOI: 10.7180/kmj.22.105.

[43] S. Keshav. „How to Read a Paper“. In: *SIGCOMM Comput. Commun. Rev.* 37.3 (Juli 2007), S. 83–84. ISSN: 0146-4833. DOI: 10.1145/1273445.1273458. URL: https://doi.org/10.1145/1273445.1273458.

[44] M. M. Kessler. „Background to scientific communication“. In: *IRE Transactions on Engineering Writing and Speech* 3.1 (1960), S. 3–6. DOI: 10.1109/TEWS.1960.6590860.

[45] Barbara Kitchenham und Stuart Charters. „Guidelines for performing Systematic Literature Reviews in Software Engineering“. In: 2 (Jan. 2007).

[46] Barbara Kitchenham u. a. „Evaluating guidelines for reporting empirical software engineering studies“. In: *Empirical Software Engineering* 13 (2008), S. 97–121.

[47] Barbara A Kitchenham u. a. „Preliminary guidelines for empirical research in software engineering“. In: *IEEE Transactions on software engineering* 28.8 (2002), S. 721–734.

[48] Masanari Kondo u. a. „An empirical study of issue-link algorithms: Which issue-link algorithms should we use?“ In: *Empirical Software Engineering* 27.6 (2022). DOI: 10.1007/s10664-022-10120-x.

[49] Stephan Kornmesser und Wilhelm Büttemeyer. *Wissenschaftstheorie: Eine Einführung*. J.B. Metzler Verlag, 2020.

[50] Thomas S. Kuhn. *The structure of Scientific Revolutions*. University of Chicago Press, 1970.

[51] Peter LaPlaca, Adam Lindgreen und Joëlle Vanhamme. „How to write really good articles for premier academic journals“. In: *Industrial Marketing Management* 68 (2018), S. 202–209. DOI: 10.1016/j.indmarman.2017.11.014.

[52] Harold N. Lee. „Scientific Method and Knowledge“. In: *Philosophy of Science* 10.2 (1943), S. 67–74. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/184289 (besucht am 04. 09. 2023).

[53] Adam Lindgreen und C. Anthony Di Benedetto. „How authors really frame a top manuscript“. In: *Industrial Marketing Management* 94 (2021), A11–A17. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2020.04.004.

[54] Giancarlo Maria Liumbruno und Claudio Velati. „How to write a scientific manuscript for publication“. In: *Blood Transfusion* (2013). ISSN: 1723-2007. DOI: 10.2450/2012.0247-12. URL: https://doi.org/10.2450/2012.0247-12.

[55] Holger Lyre. *nbsp;Interview: Wissenschaftsphilosophie: Von der Wissenschaftstheorie zur Wissenschaftsphilosophie*. März 2008. URL: https://www.information-philosophie.de/?a=1&t=939&n=2&y=4&c=68.

[56] Yuxing Ma u. a. „A methodology for analyzing uptake of software technologies among developers“. In: *IEEE Transactions on Software Engineering* 48.2 (2022), S. 485–501. DOI: 10.1109/tse.2020.2993758.

[57] L. Mateu Arrom u. a. „How to write an original article“. In: *Actas Urológicas Españolas (English Edition)* 42.9 (2018), S. 545–550. ISSN: 2173-5786. DOI: https://doi.org/10.1016/j.acuroe.2018.02.012.

[58] *Metascience*. URL: https://www.springer.com/journal/11016 (besucht am 29. 10. 2023).

[59] Michela Montesi und Patricia Lago. „Software engineering article types: An analysis of the literature“. In: *Journal of Systems and Software* 81.10 (2008), S. 1694–1714.

[60] Reza Nadri, Gema Rodriguez-Perez und Meiyappan Nagappan. „On the relationship between the developer's perceptible race and ethnicity and the evaluation of contributions in OSS“. In: *IEEE Transactions on Software Engineering* 48.8 (2022), S. 2955–2968. DOI: 10.1109/tse.2021.3073773.

[61] Sanida Omerovic u. a. „Methodology for Written and Oral Presentation of Research Results“. In: *Journal of Professional Issues in Engineering Education and Practice* 136.2 (2010), S. 112–117. DOI: 10.1061/(ASCE)EI.1943-5541.0000010.

[62] *Perspectives on Science*. URL: https://direct.mit.edu/posc (besucht am 28. 10. 2023).

[63] *Philosophy of Science*. URL: https://www.jstor.org/journal/philscie (besucht am 28. 10. 2023).

[64] Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. routledge, 2014.

[65] Julian Aron Prenner und Romain Robbes. „Making the most of small software engineering datasets with modern machine learning“. In: *IEEE Transactions on Software Engineering* (2022), S. 1–1. DOI: 10.1109/tse.2021.3135465.

[66] Georges Aaron Randrianaina u. a. „On the benefits and limits of incremental build of software configurations“. In: *Proceedings of the 44th International Conference on Software Engineering* (2022). DOI: 10.1145/3510003.3510190.

[67] John B. Reubens. „Challenge: Good Scientific Writing“. In: *IEEE Transactions on Engineering Writing and Speech* 8.2 (1965), S. 48–55. DOI: 10.1109/TEWS.1965.4322594.

[68] Per Runeson und Martin Höst. „Guidelines for conducting and reporting case study research in software engineering“. In: *Empirical software engineering* 14.2 (2009), S. 131–164.

[69] Michael Ruse. *Science and spirituality: Making room for faith in the age of science*. Cambridge University Press, 2010.

[70] Jutta Schickore. „Doing Science, Writing Science“. In: *Philosophy of Science* 75.3 (2008), S. 323–343. DOI: 10.1086/592951.

[71] *Science Philosophy*. URL: http://eiris.it/ojs/index.php/scienceandphilosophy (besucht am 28. 10. 2023).

[72] Carolyn B. Seaman. „Qualitative methods in empirical studies of software engineering“. In: *IEEE Transactions on software engineering* 25.4 (1999), S. 557–572.

[73] Forrest Shull, Janice Singer und Dag IK Sjøberg. *Guide to advanced empirical software engineering*. Springer, 2007.

[74] Shaminder Singh und Andrew Estefan. „Selecting a Grounded Theory Approach for Nursing Research“. In: *Global Qualitative Nursing Research* 5 (Okt. 2018), S. 233339361879957. DOI: 10.1177/2333393618799571.

[75] Hossein Soleimani und David J. Miller. „Semi-Supervised Multi-Label Topic Models for Document Classification and Sentence Labeling“. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM '16. Indianapolis, Indiana, USA: Association for Computing Machinery, 2016, S. 105–114. ISBN: 9781450340731. DOI: 10.1145/2983323.2983752. URL: https://doi.org/10.1145/2983323.2983752.

[76] L. Solymar. „The motivation and technique of writing scientific contributions“. In: *Proceedings of the IEEE* 51.4 (1963), S. 628–629. DOI: 10.1109/PROC.1963.2240.

[77] Ningyuan Song u. a. „Argument Structure Mining in Scientific Articles: A Comparative Analysis“. In: *Proceedings of the 18th Joint Conference on Digital Libraries*. JCDL '19. Champaign, Illinois: IEEE Press, 2020, S. 339–340. ISBN: 9781728115474. DOI: 10.1109/JCDL.2019.00060. URL: https://doi.org/10.1109/JCDL.2019.00060.

[78] Wolfgang Spohn. *nbsp;Interview: Wissenschaftsphilosophie: Von der Wissenschaftstheorie zur Wissenschaftsphilosophie*. März 2008. URL: https://www.information-philosophie.de/?a=1&t=939&n=2&y=4&c=68.

[79] Klaas-Jan Stol, Paul Ralph und Brian Fitzgerald. „Grounded theory in software engineering research: a critical review and guidelines“. In: *Proceedings of the 38th International conference on software engineering*. 2016, S. 120–131.

[80] RV Subramanyam. „Art of reading a journal article: Methodically and effectively“. In: *Journal of oral and maxillofacial pathology: JOMFP* 17.1 (2013), S. 65.

[81] Frederick Suppe. „The structure of a scientific paper“. In: *Philosophy of Science* 65.3 (Sep. 1998), S. 381–405. DOI: 10.1086/392651.

[82] *Synthese*. URL: https://www.springer.com/journal/11229 (besucht am 28. 10. 2023).

[83] *The British Journal for the Philosophy of Science*. URL: https://www.journals.uchicago.edu/toc/bjps/current (besucht am 29. 10. 2023).

[84] umnLibraries. *Tutorial: How to Read and Comprehend Scientific Research Articles*. Youtube. 2015. URL: https://www.youtube.com/watch?v=t2K6mJkSWoA#t=1m11s.

[85] Karsten Weber. „Erkenntniswege in der Wissenschaft“. In: Jan. 2012. ISBN: 978-3865681768.

[86] *What is a sprint retrospective? - Scrum.org*. URL: https://www.scrum.org/resources/what-is-a-sprint-retrospective (besucht am 28. 10. 2023).

[87] G. M. Whitesides. „Whitesides' Group: Writing a Paper“. In: *Advanced Materials* 16.15 (Aug. 2004), S. 1375–1377. DOI: 10.1002/adma.200400767. URL: https://doi.org/10.1002/adma.200400767.

[88] Claes Wohlin und Martin Höst. „Controlled experiments in software engineering“. In: *Information and Software Technology* 43.15 (2001), S. 921–924.

[89] Juliane Wundermann u. a. *How (Not) To Write a Software Engineering Conclusion: A Preliminary Report*. 2023. URL: https://git.imp.fu-berlin.de/ailis/ma-thesis-quality-of-conclusions/-/blob/main/Bericht_der_Studierenden.pdf?ref_type=heads.

A Codes zu Beginn der offenen Codierungsphase

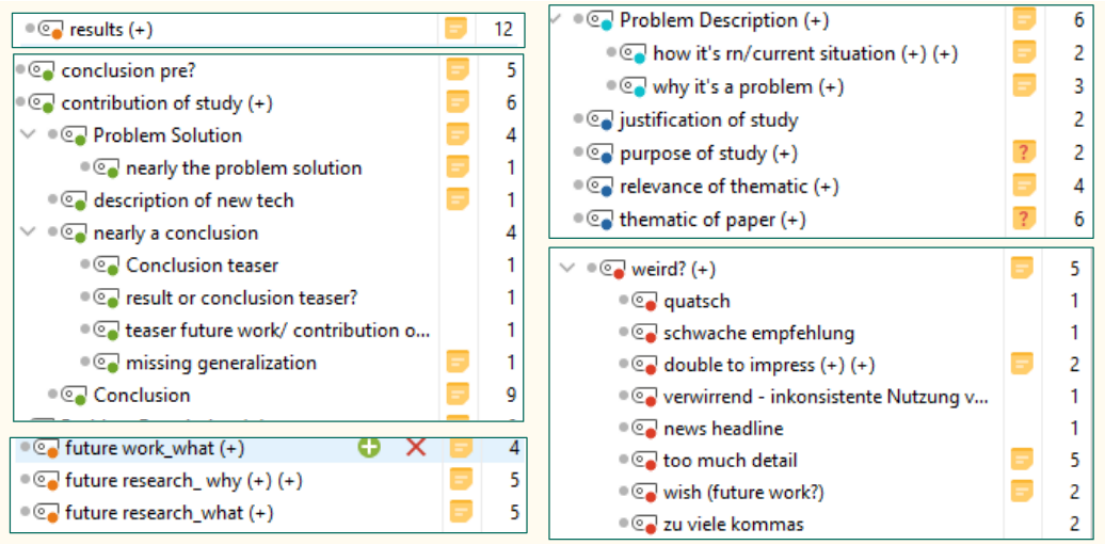

Abbildung 5: Codes nach einer ersten offenen Codierungsphase inkl. aufräumen von Codes, nach ca. 2 Wochen codieren

B Aufgabenstellungen für die offene Codierungsphase der Studierenden:

(Detailinfor zu Erprobung des Codebuchs durch Studierende)

Our study has two goals:

- Characterizing how Conclusion sections are structured and how they work
- Pointing out related quality problems.

It has three phases: Open coding, content analysis, and data analysis/interpretation. The first phase aims at improving the codebook to make it more fit for reaching this goal with a minimum of work and a maximum of reliability.

Your task is to study the codebook and as many article conclusions as needed in order to point out any improvement in this direction. To make this easier, here is a list of potential topics to which you may direct your attention. Work a bit on at least two of these and no more than four. Each item includes the level of expected difficulty (L:low, M:medium, H:high); the order is roughly one of decreasing benefit.

1. H: How many typical overall „kinds“ of Conclusion exist? Which? What are the criteria for each? One approach here may be to reconstruct what purpose(s) the authors appear to have in mind for the section. (You will need to have seen many Conclusions sections to work on this one.)
2. M: How to make the discrimination of SS and OC clearer in the codebook? One approach may be to add clauses to the role descriptions. Another may be to split some of the content codes such that they become specific to one role.
3. M: How to make the discrimination of OC and OL clearer in the codebook?

4. L: conclusion-gen annotations should have mandatory suffixes which indicate the confidence with which a conclusion is stated, e.g. conclusion-gen:medium. How many levels should we use? What are the criteria for each?
5. M: What other suffixes to the conclusion codes could add helpful information? Spell out their values and corresponding criteria.
6. H: In which cases do content codes occur in an unexpected role *without* that being quality deficiency?
7. L: How to discriminate conclusion-gen clearly from fshould?
8. M: Should we discriminate whether a conclusion pertains more to researchers or more to practitioners? Why? And how to diagnose? (This could become another suffix, e.g. conclusion-gen:practitioner)
9. M: How can we decide when a sentence that is not a conclusion should have been one? Which types of such cases exist? One type is probably the use of past tense instead of present tense?
10. M: What else could we learn from coding the table of contents? Which codes should we use?
11. L: What can the sub-headings of conclusions tell us? Which codes should we use, if any?
12. H: Would a different coding granularity than per sentence be helpful? Which? Why? How to reliably recognize the segments in that case?

C Dokumente zur informellen Studie zur Überprüfung der gefundenen Qualitäts-Faktoren

Details zu der in Unterabschnitt 5.2 besprochenen informellen Studie.

C.1 Genutzte Beispiele für ganze Conclusions-Abschnitte

- **Beispiel 1 BraFriHol22:** Developers rely on their tools and environments to complete their tasks. While new and more capable tools continue to be introduced to support increasingly complex development activities, the glue between these tools has largely remained unchanged. In this work we have examined how developers complete their tasks, finding frequent misalignment between developers and their environments. By watching 17 developers perform a controlled task and 10 industrial developers perform their own tasks, we observed developers frequently decomposing their tasks into high-level goals and then operationalizing those goals as sequences of low-level actions that they could then manually perform. The misalignment between the high-level goals the developers want to perform and the low-level actions provided by their environments induces friction that impedes their progress. We identify seven specific ways these frictions impact developers and provide design recommendations that could improve developer workflows and make them more productive.

- **Beispiel 2 - HatNovBal22** We conducted the first comprehensive mixed-methods study on how and why software developers adopt GitHub Discussions and how they perceive the new feature's usefulness. We further studied how Discussions compare to existing channels such as GitHub Issues, Pull Requests, and Stack Overflow threads. Based on an analysis of Discussions in 92 GitHub projects, a survey with developers already using the feature, and a comparison with related Stack Overflow posts, we derive the following recommendations for project managers. (i) Set guidelines for participating in discussions, as users sometimes find it difficult to choose an appropriate channel from Discussions, Issues, and Pull Requests. (ii) Encour-age core developers to participate in discussions, as it is evident that the involvement of core developers is crucial in the successful use of GitHub Discussions.

- **Beispiel 3 - MarSen22:** In this article, we discuss our experience of co-creation, as means to IA collaboration, gained in three collaboration setups in the area of SE research. Throughout this experience, we have observed a set of 28 recurring best practices and issues to avoid, which we provide as the reported patterns and anti-patterns. We exemplify the patterns and anti-patterns using three different IA collaboration projects. Such exemplified insights into recurring patterns of successes and failures can positively contribute to other IA collaboration projects in SE.

C.2 Genutzte Beispiele für einzelne Schlussfolgerungssätze

- **BalRal22 - 1, Variante A:** A novel exploration of the arguments for representativeness (see Table 2), clarifying that randomness neither equals nor guarantees representativeness, and that a nonprobability sample can be more representative than a probability sample.

  - **Varainte B:** A novel exploration of the arguments for representativeness, giving us insights to the relationship between randomness and representativeness.

- **MarSen22 -1, Variante A:** Throughout this experience, we have observed a set of 28 recurring best practices and issues to avoid, which we provide as the reported patterns and anti-patterns.

  - **Varainte B:** Throughout this experience, we have observed a set of 28 recurring best practices and issues to avoid, which we provide as the reported patterns and anti-patterns in Section 7.3. The most common patterns and anti patterns are (...).

- **PreRob22 - 2, Variante A:** In general, we advise against pre-training a new model from scratch as it is extremely resource intensive, for mixed results.

  - **Varainte B:** Pre-training a new model from scratch is extremly resource, intensive for mixed results.

- **Hat Nov Bal - 2 - 3, Variante A** Based on an analysis of Discussions in 92 GitHub projects, a survey with developers already using the feature, and a comparison with related Stack Overflow posts, we derive the following recommendations for project man-agers. ((i)... (ii)...).
  - **Varainte B:** We did an analysis of Discussions in 92 GitHub projects, a survey with developers already using the feature, and a comparison with related Stack Overflow posts. Taking the results together leads to following: ((i) ... (ii) . . .). These findings can be directly adopted by project managers.
- **RanTërKhe22 - 3, Variante A**: Our results suggest that incremental build of configurations can reduce build time without trading correctness. Developers and maintainers can already benefit from this simple approach.
  - **Varainte B:** We recommend to use an incremental build for configurations to reduce build time without lowering correctness
- **NadRodNag-4, Variante A:** Therefore, these results may indicate that OSS communities need to be aware of the diversity of their developers and move towards a more racially and ethnically diverse OSS community.
  - **Varainte B:** Therefore, these results indicate that OSS com-munities need to be aware of the diversity of their developers and move towards a more racially and ethni-cally diverse OSS community.
- **MaMocZar-4, Variante A:** Developers, at least in the context of technical decisions regarding which technology to use, do not appear to be immune from these biases.
  - **Varainte B:** Developers, at least in the context of technical decisions regarding which technology to use are not immune from these biases.
- **MaMocZar22-5, Variante A:** Integrating software supply chain concepts and models to operationalize key variables from social contagion theory to investigate software technology adoption appears to have provided a number of potentially useful insights in the pres-ent case study of two data manipulation technologies within R language.
  - **Varainte B:** To investigate software technology adoption we have integrated software supply chain concepts and models to operationalize key variables from social contagion theory. This appears to have provided a number of potentially useful insights in the pres-ent case study of two data manipulation technologies within R language.
- **ArdBotCop22-2, Variante A:** The performance and, ultimately, the applicability of the VGT paradigm strongly depends on the robustness of the underlying image recognition algorithms.
  - **Varainte B:** We have found that the performance and, ultimately, the applicability of the VGT paradigm strongly depends on the robustness of the underlying image recognition algorithms

C.3 Vorbereitetes Dokument zum festhalten von Notizen

Qualität von Conclusions-Abschnitten in der Informatik

komplette Conclusion-Abschnitte Intro: Ziel Erkenntnisse teilen, Take-Aways geben, Reminder Leser hat min. Abstract gelesen, manchmal ganzen Artikel, Leser überzeugen

- 1:BraFriHol — viel besser ○ besser ○ eher besser ○ gleich ○
- 2:HatNovBal — viel besser ○ besser ○ eher besser ○ gleich ○
- 3:MarSen — viel besser ○ besser ○ eher besser ○ gleich ○

einzelne Textbeispiele von Schlussfolgerungssätzen

- **BalRal22-1**
- **A>B**

viel besser ○ besser ○ eher besser ○ gleich ○

- **MarSen22-1**
- **B>A**

viel besser ○ besser ○ eher besser ○ gleich ○

- **PreRob22-2**
- **A>B**

viel besser ○ besser ○ eher besser ○ gleich ○

	viel besser	besser	eher besser	gleich
• **Hat NovBal-2-3** • **A>B**	○	○	○	○
• **RanTërKhe-3** • **B>A**	○	○	○	○
• **NadRodNag-4** • **B>A**	○	○	○	○
• **MaMocZar-4** • **B>A**	○	○	○	○
• **MaMocZar-5** • **B>A**	○	○	○	○
• **ArdBotCop-2** • **B>A**	○	○	○	○

1 - Teaser
2 - Eindeutige Signale für Schlussfolgerungen, **3 - Füllsätze,**
4 - squishy wording, **5- komplizierter satzbau**

C.4 detaillierte Ergebnisse

Die folgende Tabelle enthält die Antworten der Teilnehmenden. Jedes der hier bewerteten Beispiele bestand aus zwei Varianten desselben Satzes. Die Ziffern hinter dem Kürzel des Beispiels lassen darauf schließen für welchen der Faktoren das Beispiel ist, nach der Nummerierung der Auflistung oben. „MaMocZar-4“ mit den Sätzen „Developers, at least in the context of technical decisions regarding which technology to use, do not appear to be immune from these biases.“ und „Developers, at least in the context of technical decisions regarding which technology to use are not immune from these biases.“ ist Beispielsweise ein Beispiel für den Faktor *Squishy wording*. Selbst aus dem Satz ohne Kontext lässt sich schätzen, dass in der Studie zumindest ein Existenzbeweise gefunden wurde, so dass „appear to be“ einfach nur eine vorsichtigere Formulierung ist, die aber auch weniger Aussagekraft hat und dadurch von betroffenen möglicherweise nicht als relevant wahrgenommen werden würde. Da ich die finale Skala erst ab dem dritten Durchlauf hatte, habe ich die gegebene Bewertungen der ersten Teilnehmenden in die neue Skala gemappt. Dafür unterscheide ich dazwischen wie viele Skala Schritte Unterschied zwischen den Bewertungen beider Varianten liegen. Die gemappten Bewertugnen sind eingeklammert.

	Bessere Variante ist:		Bewertung des Unterschieds			
Beispiel	Erwartet	Antwort	viel besser	besser	eher besser	gleich
BalRal-1	A	A(5)	1+(1)	2	(1)	
		B(3)	1	2		
MarSen-1	B	A(0)				1
		B(7)	1+(1)	2+(1)	2	
PreRob-2	A	A(3)		3		1
		B(4)		2+(1)	(1)	
ArdBotCop-2	B	A(1)	1			1+(1)
		B(5)		3	1+(1)	
HatNovBal-2-3	A	A(5)	2	2	(1)	1
		B(2)		1	(1)	
RanTerKhe-3	B	A(6)		2	3+(1)	
		B(2)		1	(1)	
NadRodNag-4	B	A(0)				1
		B(7)	2+(1)	3+(1)		
MaMocZar-4	B	A(1)		1		1+(1)
		B(5)	1	2+(1)	1	
MaMocZar-5	B	A(0)				(1)
		B(7)	2	2+(1)	2	

Tabelle 1: Ergebnisse Schlussfolgerungssätze: Die Tabelle fasst die Antworten der Teilenhmenden Zusammen.

Ergänzend folgt eine Aufzählung der gegebenen Gründe für die jeweilige Bewertung der Sätze, beginnend mit meiner eigenen Begründung, der Anzahl an Teilnehmenden die ähnliche Begründungen zur erwarteten gegeben haben, gefolgt von den abweichenden Begründungen der Teilnehmenden.

Eine teilnehmende Person hatte zu der Gesamtheit der Satzbeispiel zu sagen, dass diese eher Resultaten entsprechen und nicht Schlussfolgerungen. Damit meinte die Person, dass einige der Schlussfolgerungen hier zu kleinteilig seien und nicht die passende Abstrakttionsstufe für eine Schlussfolgerung im Conclusions-Abschnitt hätten.

- **BalRal-1:** Version A ist besser, da sie mehr Informationen enthält und der genutzte Verweis zumindest genau ist. 3 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen gab es:
  - B ist kein Satz! Es fehlt ein Verb, daher wähle ich A, obwohl mir A eigentlich zu lang und kompliziert ist.
  - A ist konkreter als B.
  - A ist zu kompliziert geschrieben. (2 Teilnehmende)
  - A ist zu detailliert für einen Conclusions-Abschnitt, daher ist B besser.
- **MarSen-1:** B ist besser, weil dort die in A nur angedeuteten Patterns bzw. Anti-Patterns zumindest teilweise genannt werden. 5 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen und Anmerkungen gab es:
  - A ist präziser als B und außerdem macht man keine Rückverweise in Conclusions-Abschnitte!
  - Anmerkung zweier Teilnehmenden: Der konkrete Verweise ist unnötig hier.
- **PreRob-2:** A ist besser, weil der Satz eindeutig eine Schlussfolgerung ankündigt und diese klar von den Autoren kommt, während B so auch eine Hintergrundinformation sein könnte. 2 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen gab es:
  - A ist kürzer als B.
  - B ist eindeutig genug, da man als Leser in Conclusions-Abschnitten Schlussfolgerungen erwartet, außerdem ist B platzsparender.
  - B ist prägnant, kurz und ausdrucksstark. Das wäre ein schöner Schlusssatz.
  - A ist persönlicher, da ich dahinter eine empirische Untersuchung vermute gefällt mir das besser. B ist im Gegensatz dazu zu diktierend.
  - A klingt wie eine Meinung, in eine Conclusion gehören keine Meinungen, nur was erforscht, was herausgefunden wurde und maximal Empfehlungen mit persönlicher Schreibweise.
  - A ist eine sehr gut abstrahierte Aussage! B dagegen ist nur ein Resultat.
- **ArdBotCop-2:** B ist besser, weil es in Form einer auf die Autoren zurückzuführende Empfehlung geschrieben ist und damit eindeutig eine Schlussfolgerung ankündigt von der die Autoren offensichtlich überzeugt sind. 3 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen gab es:
  - B ist besser, da der Satz eindeutiger ist, A ist zu allgemein.

  - Nur als erster Satz in den Conclusions-Abschnitt wäre B besser, als Mittelsatz wäre A kürzer und daher besser.
  - A ist eindeutig besser, denn „we“ hat in einer wissenschaftlichen Arbeit nichts zu suchen!

- **HatNovBal-2-3:** A ist besser, da in B die Erkenntnisse der Studie neutral formuliert sind und daher ein zusätzlicher Satz benötigt wird um die Bedeutung der Erkenntnisse hervorzuheben. 2 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen gab es:
  - A ist psychologisch gesehen besser, da die Ratschläge zuletzt genannt werden und so besser im Kopf bleiben.Persönlich finde ich B's Stil schöner, auch wenn erst am Ende des Textes klar wird an wen sich die Ratschläge richten.
  - B ist besser, denn bei A geht nicht hervor ob die Analyse von den Autoren selber durchgeführt wurde oder nur die Ergebnisse übernommen wurden.
  - A ist kürzer und kompakter als B.
  - A ist konkreter als B, aber mich stört die persönliche Sprache.
  - A und B unterscheiden sich nur darin, dass sie aktiv bzw. passiv formuliert sind. Meine persönliche Vorliebe ist die passive Schreibweise, ich weiß aber, dass andere Forschende die aktive bevorzugen. Daher schätze ich die Sätze als gleich ein.
  - A ist vager formuliert, B ist zu übvermütig mit dem letzten Satz.

- **RanTerKhe-3:** B ist besser, da es den Inhalt von A kompakter zusammenfasst und keinen extra Füllsatz benötigt um die Empfehlung überzeugend zu machen. 1 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen gab es:
  - A ist direkter und daher schöner.
  - Bei A wird die Zielgruppe klar genannt, bei B bleibt dies unklar.
  - B ist zu unspezifisch, es bleibt unklar für wen die Empfehlung ist und wofür man dieser folgen sollte.
  - Autoren scheinen sich nicht zu trauen ihre eigenen Ergebnisse zu benennen. A wirkt unüberzeugt auf mich.
  - A ist besser weil es eine Begründung gibt „devs can beenfit“, nicht nur eine Empfehlung
  - A ist besser weil es deutlich macht, dass sich die Ergebnisse sich aus den Ergebnissen direkt ableiten.

- **NadRodNag-4:** B ist der Ausdruckstärkere klarere Satz, da er eine Aussage macht und nicht eine Vermutung aufstellt durch das „may“, wie Satz A. 6 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen bzw Anmerkungen gab es:

  - Anmerkung: Entweder ich habe was festgestellt oder nicht und das sollte ich dann auch so schreiben, aber nicht feige sein und sich vor der Diskussion drücken.
  - Je nach Resultaten ist einer der beiden Sätze angebracht und der andere nicht, daher sind diese hier für mich gleich gut.
  - Autoren scheinen sich nicht zu trauen ihre eigenen Ergebnisse zu benennen.
  - Mit „indicate“ wird ja schon eine Abstraktion der Daten durch die forscher signalisiert, das muss nicht noch durch „may“ passieren, das verdoppelt das praktisch.

- **MaMocZar-4:** B ist Ausdrucksstärker und klarer als Satz A, da er seine Aussage nicht verweichlicht. 4 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen gab es:
  - B ist besser, obwohl es weniger vorsichtig formuliert ist.
  - Sehr ähnliche Sätze, einer ist politisch Korrekter, der andere ist klarer, daher sind beide gleich gut.
  - Macht für mich keinen Unterschied.
  - Autoren scheinen sich nicht zu trauen ihre eigenen Ergebnisse zu benennen.
  - In Satz A wird zu viel drumherumgelabert, die Füllwörter nerven.
  - Satz A ist besser, B ist zu absolut formuliert und hat eine schlechte Kommasetzung.

- **MaMocZar-5:** B ist einfacher zu lesen, da er aus zwei Sätzen besteht, anstatt wie A aus einem sehr langen Satz der dazu führt, dass Lesende diesen Satz öfter als einmal lesen müssen. 7 Teilnehmende lieferten ähnliche Begründungen. Folgende abweichende Begründungen und Anmerkungen gab es:
  - Beide Varianten sind gleich gut. Mich persönlich stören lange Sätze nicht, ich bin das geübt.
  - Anmerkung: Wäre ich geübter im Englischen, wäre ich für den längeren Satz!