IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

Verbundprojekt

TextTransfer (Pilot)

Korpusgestützte Erkennung von Verwertungsmustern in wissenschaftlichen Texten

Abschlussbericht Gesamtprojekt

nach Nr. 3.2. BNBest-BMBF 98

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

DLR Projektträger

1. Eckdaten des Forschungsprojektes

Vorhabenbezeichnung	*TextTransfer (Pilot)* - Korpusgestützte Erkennung von Verwertungsmustern in wissenschaftlichen Texten
Projektart	Verbundprojekt
Verbundpartner	▪ Leibniz-Institut für Deutsche Sprache (IDS), Mannheim *Teilprojekt IDS: Analysemethoden und Anwendungsfälle* ▪ Technische Informationsbibliothek (TIB) - Leibniz-Informationszentrum Technik und Naturwissenschaft, Universitätsbibliothek, Hannover *Teilprojekt TIB: Datenmanagement, Dokumentenauswahl sowie Workflow zur Bereitstellung eines XML-basierten Korpus*
Zuwendungsempfänger und ausführende Stellen	▪ Leibniz-Institut für Deutsche Sprache (IDS), R5, 6-13, 68161 Mannheim ▪ Technische Informationsbibliothek (TIB) - Leibniz-Informationszentrum Technik und Naturwissenschaft, Universitätsbibliothek, Welfengarten 1B, 30167 Hannover
Gesamtprojektleitung	Prof. Dr. Andreas Witt (IDS) (witt@ids-mannheim.de)
Förderkennzeichen	▪ 01IO1634 (IDS) ▪ 01IO1635 (TIB)
Förderer	Bundesministerium für Bildung und Forschung (BMBF)
Projektträger	Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Bereich Gesellschaft, Innovation, Technologie, Heinrich-Konen-Straße 1, 53227 Bonn
Laufzeit des Vorhabens	01.12.2016 – 31.12.2019 (inkl. Kostenneutraler Laufzeitverlängerung)
Berichtszeitraum	1.12.2016 – 31.12.2019
Berichtstyp	Gesamtabschlussbericht[1]
Stichwörter	Text Mining, Maschinelles Lernen, Korpusanalyse, Computerlinguistik, Korpuslinguistik, Impact, Impact Assessment, Wissenstransfer, Forschungsimpact, Impact-

[1] Nach Absprache zwischen DLR und IDS ist auf Grund des zeitlichen Zusammenfallens zwischen Jahresbericht 2019 und Ende des Gesamtprojektes zum Ende 2019 kein gesonderter Jahresbericht 2019 erforderlich (vgl. hierzu u.a. Emailkommunikation vom 18.11.2019.)

	Indikatoren, Transfer-Potenzial, IDS, TIB
Version	1.0
Verfasser	Prof. Dr. Andreas Witt (IDS)
Datum	16.06.2020

Inhalt:

FKZ 01IO1634
Version 1.0

Abbildungsverzeichnis:

2. Kurze Darstellung

2.1. Aufgabenstellung

Die zentrale Aufgabenstellung des Verbundprojektes *TextTransfer (Pilot)* war eine Machbarkeitsprüfung für die Entwicklung eines Text-Mining-Verfahrens, mit dem Forschungsergebnisse automatisiert auf Hinweise zu Transfer- und Impactpotenzialen untersucht werden können. Das vom Projektkoordinator IDS verantwortete Teilprojekt konzentrierte sich dabei auf die Entwicklung der methodischen Grundlagen, während der Projektpartner TIB vornehmlich für die Bereitstellung eines geeigneten Datensatzes verantwortlich war. Solchen automatisierten Verfahren liegen zumeist textbasierte Daten als physisches Manifest wissenschaftlicher Erkenntnisse zugrunde, die im Falle von *TextTransfer (Pilot)* als empirische Grundlage herangezogen wurden. Das im Verbund zur Anwendung gebrachte maschinelle Lernverfahren stützte sich ausschließlich auf deutschsprachige Projektendberichte öffentlich geförderter Forschung. Diese Textgattung eignet sich insbesondere hinsichtlich ihrer öffentlichen Verfügbarkeit bei zuständigen Gedächtnisorganisationen und aufgrund ihrer im Vergleich zu anderen Formaten wissenschaftlicher Publikation relativen strukturellen wie sprachlichen Homogenität. *TextTransfer (Pilot)* ging daher grundsätzlich von der Annahme struktureller bzw. sprachlicher Ähnlichkeit in Berichtstexten aus, bei denen der Nachweis tatsächlich erfolgten Transfers zu erbringen war. Im Folgenden wird in diesen Fällen von Texten bzw. textgebundenen Forschungsergebnissen mit Transfer- und Impactpotenzial gesprochen werden. Es wurde ferner postuliert, dass sich diese Indizien von sprachlichen Eigenschaften in Texten zu Projekten ohne nachzuweisenden bzw. ggf. auch niemals erfolgtem, aber potenziell möglichem Transfer oder Impact unterscheiden lassen. Mit einer Verifizierung dieser Annahmen war es möglich, Transfer- oder Impactwahrscheinlichkeiten in großen Mengen von Berichtsdaten ohne eingehende Lektüre zu prognostizieren.

Nach Festlegung des Bezugsrahmens, im Rahmen dessen u.a. der für die Fragestellung erfolgversprechenden Dokumententyp als auch spezifische thematisch-fachliche Domänen definiert und eine Grundgesamtheit an Projektberichten generiert wurde, war für eine geeignete, maschinenlesbare Datenbasis eine in das maschinenlesbare Format TEI XML (i5) konvertierte

FKZ 01IO1634
Version 1.0

Stichprobe zu ziehen, die verwertbare Ergebnisse zu produzieren versprach. Der Projektansatz bestand darin, für das überwachte Lernverfahren (supervised machine learning) zunächst ein Trainingsdatenset aus der Stichprobe auszuwählen, das mit bestimmten impactrelevante Texteigenschaften repräsentierenden Informationen (Klassifikation) angereichert wurde. Somit konnte die Maschine auf vorgegebene Zusammenhänge von Texteigenschaften und Impact-Klassifikationen trainiert werden. In einem zweiten Schritt wurden dann Indizien für ähnliche Zusammenhänge in der Maschine unbekannten, nicht vorab klassifizierten Textmengen (Evaluationsset) gesucht (distant reading).

Eine zentrale Fragestellung des Projektes war dabei, welche Informationen über den tatsächlichen Transfer und Impact einzelner Forschungsergebnisse in welcher Form gewonnen und den Projektergebnissen zugeordnet werden können.

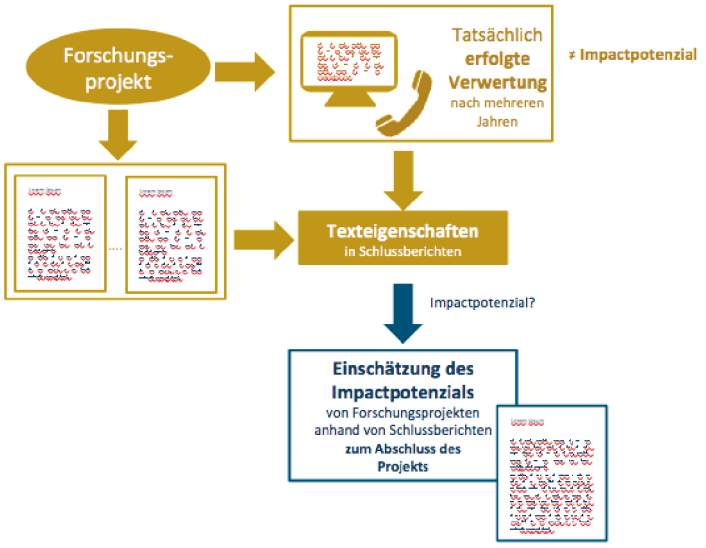

Abb. 1 Untersuchungsgegenstand TextTransfer

Zu identifizieren waren somit spezifische Texteigenschaften, die ein Transfer- und Impactpotenzial von Forschungsergebnissen nahelegen. Die Suchmethode musste außerdem in der Lage sein, sehr

FKZ 01IO1634
Version 1.0

große Datenmengen, wie sie sich durch zahlreiche Projektanträge und -berichte bei den zuständigen Fachbibliotheken anhäufen, gezielt nach bestimmten transfer- und impactrelevanten Themen mit möglichst wenig Aufwand zu durchsuchen.

Ein automatisiertes Verfahren war zu entwickeln, das in erster Linie die Wissenschaft unterstützt, Transfer- und Impactpotenziale in wissenschaftlichen Texten besser zu identifizieren und so den Wirkungsgrad von Investitionen in die Forschung zu optimieren. Im Ergebnis lag eine auf maschinellem Lernen basierende Methode vor, die mittels statistischem, textstrukturellem Vergleich sprachlicher Ähnlichkeiten in Projektendberichten automatisiert Transfer- und Impact-Wahrscheinlichkeiten nachzuweisen in der Lage ist. Sowohl mit Blick auf diese Fertigkeit als auch auf die eigens erstellte, komplexe Datengrundlage stellt das Vorhaben ein Unikat sowie ein Novum dar. Insbesondere gilt dies für die Klassifikation hochwertig angereicherter, deutschsprachiger Datensätze, wie sie bisherige auf Spracherkennung trainierte Verfahren – zu nennen wären etwa gängige Suchmaschinen – nicht zu leisten vermögen.

2.2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Voraussetzung für das Projekt waren die Kompetenzen, die die Projektpartner jeweils in das Projekt mit einbrachten. Für die Bearbeitung deutschsprachiger Textdaten hinsichtlich der im Projekt gewählten Fragestellung bestanden keine etablierten Strukturen, sodass sich ein neuer Verbund mit Kompetenzen im Bereich Korpusaufbau, Text Mining und maschinellem Impact Assessment bilden musste. Die Aufgaben der Verbundpartner stellten sich dabei wie folgt dar:

Der Projektpartner **Leibniz-Institut für Deutsche Sprache** (**IDS**[2]) konzentrierte sich im Rahmen von *TextTransfer (Pilot)* auf die Indikatoren- bzw. Kategorienschemata-Entwicklung, die Annotation von Textquellen sowie auf die exemplarische Adaption vorhandener Softwarelösungen für den Funktionsnachweis der Methode und steckte die Rahmenbedingungen für deren Anwendung ab. Das IDS hatte außerdem die Gesamtprojektleitung als auch die Projektkoordination inne.

[2] Vgl. u.a. http://www1.ids-mannheim.de

Die Aufgabe des Projektpartners **Technische Informationsbibliothek - Leibniz-Informationszentrum Technik und Naturwissenschaft, Universitätsbibliothek TIB**[3] war es, ein bedarfsgerecht zugeschnittenes, maschinenlesbares Korpus von Forschungsberichten als Stichprobe zur Verfügung zu stellen. Die TIB gewährleistete außerdem die Extraktion aus vielfältigen elektronischen Quellformaten in das Zielformat TEI XML (i5), das Textmodell des IDS[4].

Unterstützt wurden die Verbundpartner von den Unterauftragnehmern **Görgen & Köller GmbH** (**G&K)**[5] und **Prof. Dr. Jana Diesner** von der **School of Information Sciences / The *iSchool* der Universität von Illinois at Urbana-Champaign** (**UIUC**)[6] und ihrer Arbeitsgruppe. G&K unterstützte das IDS bei der Erbringung von Transfernachweisen für die die Stichprobe bildenden Projekte bzw. deren Berichte sowie ersten Überlegungen zu einer institutseigenen Anwendung der Methode selbst. UIUC passte die zur Textanalyse notwendigen Softwarelösungen durch maschinelle Lernverfahren an.

Die Aufgabenverteilung im Verbundprojekt entsprachen den Kernkompetenzen der jeweiligen Partner.

2.3. Planung und Ablauf des Vorhabens

2.3.1. Planung

Mit seinen auf Teilprojektebene angesiedelten Einzelkomponenten konnte die Vorstudie *TextTransfer (Pilot)* eine erste Basis dafür liefern, die Bedarfslücke für die maschinengestützte semantische Analyse deutschsprachiger Textdaten langfristig zu schließen. Der vergleichsweise Rückstand des deutschen Marktes in diesem Bereich ist nicht zuletzt auf personelle Engpässe qualifizierter Computerlinguistinnen/Computerlinguisten und Informatikerinnen/Informatiker im wissenschaftlichen Umfeld als auch einer Zurückhaltung in der Produktentwicklung im Bereich der Künstlichen Intelligenz auf Unternehmensseite im deutschsprachigen Raum zurückzuführen. Das Projekt wird entsprechende Fähigkeiten der semantischen Textanalyse auf Basis maschineller

[3] Vgl. u.a. https://www.tib.eu/de/
[4] Vgl. u.a. https://www1.ids-mannheim.de/kl/projekte/korpora/textmodell.html
[5] Vgl. u. a. https://gk-mb.com
[6] Vgl. u.a. https://ischool.illinois.edu/people/jana-diesner

Lernverfahren aufbauen, sollte eine Förderphase des Hauptprojektes *TextTransfer*, das sich zum Zeitpunkt der Berichtserstellung in der Antragsphase befindet, zustande kommen.

Unter diesen Voraussetzungen gestaltete sich die Projektdurchführung des in Deutschland bisher unerprobten, prototypischen Verfahrens vornehmlich auf experimenteller Basis. Der endgültige Projektansatz zur Methodenentwicklung aus dediziertem Stichprobendesign, Informationsgewinnung, Systematik der Impact-Klassifikation sowie der technischen Umsetzung im maschinellen Lernverfahren wurden erst im Zuge des Projektfortschritts finalisiert. Entsprechende Anpassungen von der ursprünglichen antragsgemäßen Umsetzungsplanung werden im Folgenden erläutert (vgl. hierzu Kap. 3, insbesondere *3.1.3 Inventar*).

Um diese Herausforderungen abzufedern, war bereits in der Antragsphase für *TextTransfer (Pilot*) auf IDS-Seite vorgesehen, die Stelle einer wissenschaftlichen Mitarbeiterin bzw. eines wissenschaftlichen Mitarbeiters mit Schwerpunkt Computerlinguistik (Pos. F082) personengebunden zu besetzen. Aufgrund unvorhergesehener personeller Änderungen konnte die Stelle nicht an die eingeplante Mitarbeiterin des IDS vergeben werden. In Abstimmung mit dem Projektträger konnte sie stattdessen zunächst nur übergangsweise und erst zum Oktober 2017 endgültig besetzt werden. Diese und weitere administrative Verzögerungen auf Seiten des Fördermittelgebers bei der Freigabe von Projektmitteln für außereuropäische Unteraufträge bzw. infolgedessen entstandene Umschichtungen im Kostenplan als auch eine damit im Zusammenhang stehende erste kostenneutrale Laufzeitverlängerung des Projektes um vier Monate (Dezember 2018 bis März 2019) konnten vom IDS in Zusammenarbeit mit dem Projektträger gelöst werden. Um die Rahmenbedingungen für eine unmittelbare Anschlussfinanzierung für eine weitere Projektphase zu schaffen, im Zuge derer die Pilotstudie zu stabilisieren und auszuweiten war, wurde auf Initiative des Projektpartners IDS eine zweite, aus Eigenmitteln der Projektpartner IDS und TIB gestemmte Phase der kostenneutralen Verlängerung von neun Monaten (April bis Dezember 2019) in die Wege geleitet.

FKZ 01IO1634
Version 1.0

Auch beim Projektpartner TIB konnte nach einigen Verzögerungen bzw. Zwischenlösungen die zu besetzende Stelle eines Informatikers (m/w/d), die u.a. die Konvertierung der Daten in das IDS Textformat vorsah, erst Anfang 2018 final besetzt werden.

Die Zielerreichung des Verbundes im Berichtszeitraum war davon unberührt.

2.3.2. Abstimmung innerhalb von *TextTransfer (Pilot)*

Die einzelnen Arbeits- und Projektschritte zwischen den Partnern IDS und TIB wurden im IDS koordiniert. Entsprechend wurden auch den Unterauftragnehmern einzelne Arbeits- und Projektschritte zugewiesen.

Zu Beginn des Projektes waren auf Grund der Klärung der Rahmenbedingungen in den einzelnen APs vermehrt physische Treffen nötig. Im weiteren Verlauf des Projektes konnten die physischen Treffen deutlich reduziert und weitgehend durch virtuelle Treffen ersetzt werden.

Statustreffen (physisch) des Projekts im Projektzeitraum:

- 17.02.2017 (G&K): Statustreffen des Projekts TT – IDS, TIB, G&K (Hürth),
- 07.03.2017 (G&K): Abstimmungsgespräch IDS – G&K (Hürth),
- 30.03.2017 (IDS): Statustreffen des Projekts TT – DLR, IDS, TIB, G&K (Mannheim),
- 13.07.2017 (G&K): Statustreffen des Projekts TT – IDS, TIB, G&K (Hürth),
- 30.07.2017 (IDS): Abstimmungsgespräch IDS – G&K (Mannheim),
- 12.09.2017 (Leibniz Gemeinschaft): Statustreffen des Projekts TT – IDS, TIB, G&K (Berlin),
- 08.11.2017 (G&K): Statustreffen des Projekts TT – IDS, TIB, G&K (Hürth),
- 21.02.2018 (IDS): Statustreffen des Projekts TT (P) – DLR, IDS, TIB, G&K, UIUC zzgl. Gast ZB MED (Mannheim),
- 23.11.2018 (IDS): Statustreffen des Projekts TT (P) – IDS, TIB, G&K, UIUC (Mannheim).

Zwecks regelmäßigem Informationsaustauschs als auch zur Abstimmung der Aktivitäten der einzelnen Partner waren neben den o.g. physischen Vor-Ort-Status-Projekttreffen jeweils am zweiten und vierten Donnerstag eines Monats bei Bedarf ein telefonisches Jour Fixe vorgesehen, an dem

jeweils mindestens eine Mitarbeiterin bzw. ein Mitarbeiter eines jeden Partners teilnehmen sollte. Für diese Meetings wurde mittels einer speziell für das Projekt angelegten Conference-Call-Nummer auf die verfügbaren Konferenz-Systeme des DFN zurückgegriffen.

Zwecks Absprachen und Zusammenarbeit fanden bei Bedarf außerdem Treffen - virtueller und nicht-virtueller Art - in kleineren Gruppen statt.

Darüber hinaus wurde der Projektstatus Quo bei für das Gesamtprojekt relevanten Entwicklungen zeitnah per Cloudlösung oder per Email an alle Projektbeteiligte kommuniziert.

Zum Austausch von Daten größeren Umfangs wurde sowohl auf Gigamove als auch auf instituts-interne Cloud-Lösungen der Projektpartner TIB und IDS zurückgegriffen.

2.3.3. Danksagung

Das Pilotprojekt *"TextTransfer - Korpusbasierte Erkennung von Verwertungsmustern in wissenschaftlichen Texten"* wurde vom Bundesministerium für Bildung und Forschung (BMBF) unter den Förderkennzahlen 01IO1634 (IDS) und 01IO1635 (TIB) gefördert und vom Deutschen Zentrum für Luft- und Raumfahrt e.V. (DLR) betreut.

Unser Dank gilt neben den Förderern allen am Projekt Beteiligten, dabei insbesondere den Unterauftragnehmern G&K und UIUC, die dank ihrer Unterstützung auch nach dem offiziellen Projektende und ihrer weiterhin sehr engagierten Mitarbeit im Rahmen der kostenneutralen Verlängerungen zu den Ergebnissen des Projektes maßgeblich mit beigetragen haben.

2.3.4. Rechte

2.4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde, insbesondere

2.4.1. - Angabe bekannter Konstruktionen, Verfahren und Schutzrechte, die für die Durchführung des Vorhabens benutzt wurden.

Für die Durchführung des Vorhabens wurden keine bekannten Konstruktionen, Verfahren oder Schutzrechte benutzt.

2.4.2. - Angabe der verwendeten Fachliteratur sowie der benutzen Informations- und Dokumentationsdienste

2.4.2.1. Fachliteratur

Aksnes, D. W., Langfeldt, L., and Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. SAGE Open, 9(1):2158244019829575.

Barrett, D. and Leddy, S. (2008). Assessing creative media's social impact. The Fledgling Fund.

Becker, D. R., Harris, C. C., McLaughlin, W. J., and Nielsen, E. A. (2003). A participatory approach to social impact assessment: the interactive community forum. Environmental Impact Assessment Review, 23(3):367– 382.

Becker, H. A. (2001). Social impact assessment. European Journal of Operational Research, 128(2):311–321.

Berendt, B. (2019). Ai for the common good?! pitfalls, challenges, and ethics pen-testing. Paladyn, Journal of Behavioral Robotics, 10(1):44–65.

Blakley, J., Huang, G., Nahm, S., and Shin, H. (2016). Changing appetites & changing minds: Measuring the impact of" food, inc.". The USC Annenberg Norman Lear Center.

Bornmann, L. and Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work? Scientometrics, 65(3):391–392.

FKZ 01IO1634
Version 1.0

Bornmann, L. (2012). Measuring the societal impact of research: research is less and less assessed on scientific impact aloneâwe should aim to quantify the increasingly important contributions of science to society. EMBO reports, 13(8):673–676.

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? a literature survey. Journal of the American Society for Information Science and Technology, 64(2):217–233.

Bornmann, L. (2015). Usefulness of altmetrics for measuring the broader impact of research: A case study using data from plos and f1000prime. Aslib Journal of Information Management, 67(3):305–319.

Bornmann, L. (2017). Measuring impact in research evaluations: a thorough discussion of methods for, effects of and problems with impact measurements. Higher Education, 73(5):775–787.

Chattoo, C. B. and Das, A. (2014). Assessing the social impact of issues-focused documentaries: Research methods & future considerations. Center for Media and Social Impact, School of Communication at American University.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357.

Clark, J. and Abrash, B. (2011). Social justice documentary: Designing for impact. Center for Social Media.

Diesner, J., Rezapour, R., and Jiang, M. (2016). Assessing public awareness of social justice documentary films based on news coverage versus social media. IConference 2016 Proceedings.

Gomes, D. and Stavropoulou, C. (2019). The impact generated by publicly and charity-funded research in the united kingdom: a systematic literature review. Health research policy and systems, 17(1):22.

FKZ 01IO1634
Version 1.0

Gori, M. and Pucci, A. (2006). Research paper recommender systems: A random-walk based approach. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06), pages 778–781. IEEE.

Greenhalgh, T., Raftery, J., Hanney, S., and Glover, M. (2016). Research impact: a narrative review. BMC medicine, 14(1):78.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 21(3):267–297.

Heyeres, M., Tsey, K., Yang, Y., Yan, L., and Jiang, H. (2019). The characteristics and reporting quality of research impact case studies: A systematic review. Evaluation and program planning, 73:10–23.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences, 102(46):16569–16572.

Holden, G., Rosenberg, G., and Barker, K. (2005). Bibliometrics: A potential decision making aid in hiring, reappointment, tenure and promotion decisions. Social Work in Health Care, 41(3-4):67–92.

Latané, B. (1981). The psychology of social impact. American psychologist, 36(4):343.

Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., Dempsey, E., et al. (2013). Textblob: simplified text processing; 2018.

Ma, Y. and Uzzi, B. (2018). Scientific prize network predicts who pushes the boundaries of science. Proceedings of the National Academy of Sciences, 115(50):12608– 12615.

Ma, Y., Oliveira, D. F., Woodruff, T. K., and Uzzi, B. (2019). Women who win prizes get less money and prestige.

Mishra, S., Fegley, B. D., Diesner, J., and Torvik, V. I. (2018). Self-citation is the hallmark of productive authors, of any gender. PloS one, 13(9):e0195773.

Parker, J. and Van Teijlingen, E. (2012). The research excellence framework (ref): Assessing the impact of social work research on society. Practice, 24(1):41–52.

FKZ 01IO1634
Version 1.0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. Journal of machine learning research, 12(Oct):2825–2830.

Piwowar, H. (2013). Altmetrics: Value all research products. Nature, 493(7431):159.

Pulido, C. M., Redondo-Sama, G., Sordé-Martí, T., and Flecha, R. (2018). Social impact in social media: A new method to evaluate the social impact of research. PloS one, 13(8):e0203117.

Rezapour, R. and Diesner, J. (2017). Classification and detection of micro-level impact of issue-focused documentary films based on reviews. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pages 1419–1431. ACM.

Shmueli, G. et al. (2010). To explain or to predict? Statistical science, 25(3):289–310.

Smalheiser, N. R. and Torvik, V. I. (2008). The place of literature-based discovery in contemporary scientific practice. In Literature-based discovery, pages 13–22. Springer.

Subramanyam, K. (1983). Bibliometric studies of research collaboration: A review. Journal of information Science, 6(1):33–38.

Swanson, D. R., Smalheiser, N. R., and Torvik, V. I. (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings. Journal of the American Society for Information Science and Technology, 57(11):1427–1439.

Taylor, M. (2013). Exploring the boundaries: How altmetrics can expand our vision of scholarly communication and social impact. Information Standards Quarterly, 25(2):27–32.

Tsey, K., Onnis, L.-a., Whiteside, M., McCalman, J., Williams, M., Heyeres, M., Lui, S. M. C., Klieve, H.,

Cadet-James, Y., Baird, L., et al. (2019). Assessing research impact: Australian research council criteria and the case of family wellbeing research. Evaluation and program planning, 73:176–186.

Tsey, K. (2019). Planning for and tracking research impact: Australian research council framework. In Working on Wicked Problems, pages 65–74. Springer.

Universität Koblenz-Landau, Zentrales Institut für Scientific Entrepreneurship & International Transfer, 15.08.2016: Wertschöpfender Wissens- und Technologietransfer außeruniversitärer Forschungseinrichtungen: Schlussbericht; FKZ 03 IO 1314 (https://edocs.tib.eu/files/e01fb16/871694468.pdf).

Van Raan, A. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics, 36(3):397–420.

Van Raan, A. F. (2004). Measuring science. In Handbook of quantitative science and technology research, pages 19–50. Springer.

Vanclay, F. (2006). Principles for social impact assessment: A critical comparison between the international and us documents. Environmental Impact Assessment Review, 26(1):3–14.

Verbeek, A., Debackere, K., Luwel, M., and Zimmermann, E. (2002). Measuring progress and evolution in science and technology–i: The multiple uses of bibliometric indicators. international Journal of management reviews, 4(2):179–211.

Witt, A., Diesner, J., Steffen, D., Rezapour, R., Bopp, J., Fiedler, N., Köller, C., Raster, M., and Wockenfuß, J. (2018). Impact of scientific research beyond academia: an alternative classification schema. Proceedings of the LREC 2018 Workshop on Computational Impact Detection from Text Data, pages 34–39.

Wolf, B., Lindenthal, T., Szerencsits, M., Holbrook, J. B., and Heß, J. (2013). Evaluating research beyond scientific impacthow to include criteria for productive interactions and impact on practice and society. GAIA-Ecological Perspectives for Science and Society, 22(2):104–114.

Wyndham, J., Vitullo, M., Kraska, K., Sianko, N., Carbajales, P., Nuñez-Eddy, C., and Platts, E. (2017). Giving meaning to the right to science: A global and multidisciplinary approach. Washington, DC: AAAS.

2.4.2.2. Informations- und Dokumentationsdienste

- Förderkatalog des Bundes/Projektsuche: https://foerderportal.bund.de/foekat/jsp/SucheAction.do?actionMode=searchmask
- Gemeinsamer Bibliotheksverbund: https://www.gbv.de
- TIB Informations- und Dokumentationsdatenbank: https://www.tib.eu/de/ und https://www.tib.eu/de/recherchieren-entdecken/sammelschwerpunkte/deutsche-forschungsberichte/
- Gemeinsamer Bibliotheksverbund: https://www.gbv.de

2.5. Zusammenarbeit mit anderen Stellen

Die Komplexität der Methodenentwicklung war nur durch die enge Zusammenarbeit und den jeweiligen Kernkompetenzen zwischen den Projektpartnern TIB und IDS bzw. dessen Unterauftragnehmern G&K und UIUC möglich (vgl. hierzu *Kapitel 2.2 Voraussetzungen*).

FKZ 01IO1634
Version 1.0

3. Eingehende Darstellung

3.1. Verwendung der Zuwendung und der erzielten Ergebnisse im Einzelnen

Bei Beendigung des Projektes waren die wissenschaftlich-technischen Ergebnisse wie bei Antragstellung geplant erreicht. Nachfolgend finden sich die Ergebnisse aus den Arbeitspaketen ausführlich dargestellt.

3.1.1. AP1: Bezugsrahmen

Das Arbeitspaket 1 diente übergreifend der Schaffung von Voraussetzungen, um für die einzelnen Arbeitsschritte der Verbundpartner den Projektgegenstand abgrenzen zu können.

Erste Eingrenzung: Forschungsgegenstand Projektbericht

Neben zeitintensiver, hochkomplexer Grundlagenforschung sind es hier immer wieder drittmittelfinanzierte, zeitlich umrissene und thematisch fokussierte Forschungsprojekte, die das Steuerungsinstrument der Förderer darstellen, um angesichts endlicher finanzieller Ressourcen den gesellschaftlichen Bedarf an forschungsbasiertem Wissen gezielt zu adressieren und dynamische Entwicklungen bewältigbar erscheinen zu lassen. Zum Nachweis von Legitimität der Investition durch Zielerreichung ist es üblich, dass dieses Forschungsphänomen in großer Zahl sein ureigenes Textformat hervorbringt: Projektendberichte leiten nicht nur die verfolgten Ansätze und Verfahren wissenschaftlicher Arbeit her. Auch hinsichtlich ihrer öffentlichen Verfügbarkeit bei der als Datengeber kooperierenden TIB und aufgrund ihrer im Vergleich zu anderen Formaten wissenschaftlicher Publikation relativen strukturellen und sprachlichen Homogenität (Berichtsduktus, Standardaufbau, dedizierte Transferabschnitte) wurde diese Textgattung für ein empirisch gestütztes und maschinell getriebenes Auswertungsverfahren des Textmining als besonders geeigneter Ausgangspunkt für den Funktionsnachweis erachtet.

Als erweiterter Bezugsrahmen wurden vom Verbundpartner IDS daher Forschungsprojekte ausgewählt,

- die aus öffentlich geförderten Verbundprojekten stammen;

FKZ 01IO1634
Version 1.0

- deren Projektpartner sowohl aus öffentlicher Forschung als auch der freien Wirtschaft in Deutschland kommen;
- deren deutschsprachige Abschlussberichte in der TIB frei zugänglich in digitaler Form vorliegen;
- deren Anzahl von digital verfügbaren Einzelberichten in der TIB nicht größer als zehn zum Zeitpunkt der Erhebung war;
- deren Einzelberichte keinen größeren Umfang als jeweils 350 Seiten besitzen;
- deren Projektende zwischen 2005 und 2015 liegt;
- deren Forschungsgegenstand zum Zeitpunkt der Laufzeit des Projektes *TextTransfer (Pilot)* nach subjektiver Einschätzung eine gesellschaftliche Relevanz besitzt und damit eine hohe Transferwahrscheinlichkeit.

Zweite Eingrenzung: Domäne Mobilität

Der Projektpartner TIB stellte in seinem Teilprojekt die für die Erreichung der Ziele im Verbund notwendigen digitalen Daten in Form von Projektendberichten zur Verfügung. Auf dem Weg zu einer geeigneten Datenbasis galt es daher, eine Stichprobe aus der verfügbaren Datenmenge der TIB zu ziehen, von der mit Blick auf den gewählten Ansatz eine größtmögliche Erfolgswahrscheinlichkeit zu erwarten war. Um durch fachsprachliche Effekte ausgelöste Verzerrungen möglichst abzumildern und sprachliche wie strukturelle Vergleichbarkeit der Texte herzustellen, wurde der Forschungsgegenstand zunächst *thematisch* stark eingegrenzt – perspektivisch wird in einer zweiten Förderphase zu prüfen sein, inwiefern sich auf Transfer und Verwertung hinweisende Textelemente losgelöst von einem prägenden „Fachjargon“ disziplinübergreifend ähneln und hinsichtlich ihres Transfer- und Impactpotenzials vergleichbar sind.

Um den Forschungsgegenstand weiter einzugrenzen, fiel die Entscheidung auf Forschungsprojekte der Domäne **Mobilität,**

- die von unterschiedlichen Projektträgern gefördert wurden;
- die unterschiedliche Dimensionen des Themas Mobilität repräsentierten (Elektronik, autonomes elektronisches Fahren, Elektromobilität, Mensch-Technik-Interaktion, Digitalisierung, Industrie 4.0, Entwicklung digitaler Technologien, demographischer Wandel);
- die das Förderprofil „Technologie und Innovationsförderung“ besaßen.

FKZ 01IO1634
Version 1.0

Insgesamt erfüllten 267 Projekte mit circa 1.000 Einzelberichten die genannten Kriterien und stellten damit die Grundgesamtheit als Basis für die weitere Bearbeitung dar.

Mit Ausnahme der Kriterien

- gesellschaftliche Relevanz
- hohe Wahrscheinlichkeit der Transferrelevanz des Themas
- Durchführbarkeit

konnte die Ausgestaltung der Grundgesamtheit in ihrer vorliegenden Art als willkürlich bezeichnet werden, da die Auswahl der Kriterien auf Metaebene für die Entwicklung einer Methodologie nicht entscheidend war. Die relevanten Berichte wurden dem Projektpartner TIB in entsprechend den gemeinsam entwickelten Spezifika aufbereiteter Form zur Verfügung gestellt. Eine geeignete Pipeline wurde hierfür entwickelt.

Für die maschinelle Analyse wurde die Wirksamkeit von Annotationen mit transferbezogenen Informationen der Quellenbasis geprüft. Hierfür war die Gewinnung von Transfer- bzw. Impactnachweisen für Projekte und deren repräsentative Berichte notwendig. Generierte die trainierte Maschine im distant reading konkrete Prognosen anhand nichtannotierter Texte, mussten diese mittels dem Analyseverfahren vorenthaltenen Informationen abgeglichen werden, um den Funktionsnachweis der Methode zu erbringen. Da sich solche Transfernachweise auf unterschiedlichen Ebenen der Manifestation von Forschungsergebnissen finden lassen, war – wie in Abbildung 2 dargestellt -– eine weitere Rahmenbedingung die Festlegung der Betrachtung der jeweiligen Projekte unter zwei Blickwinkeln: Projekte auf Berichts-/Textebene (vgl. hierzu auch im Folgenden unter AP 5 zum Thema: Induktiver, Bottom-Up-Ansatz: Von den Daten zur Theorie) und Projekte auf (externer) Realitätsebene (Projektebene - vgl. hierzu auch im Folgenden unter AP 5 zum Thema: Deduktiver, Top-Down-Ansatz: Von der Theorie zu den Daten).

FKZ 01IO1634
Version 1.0

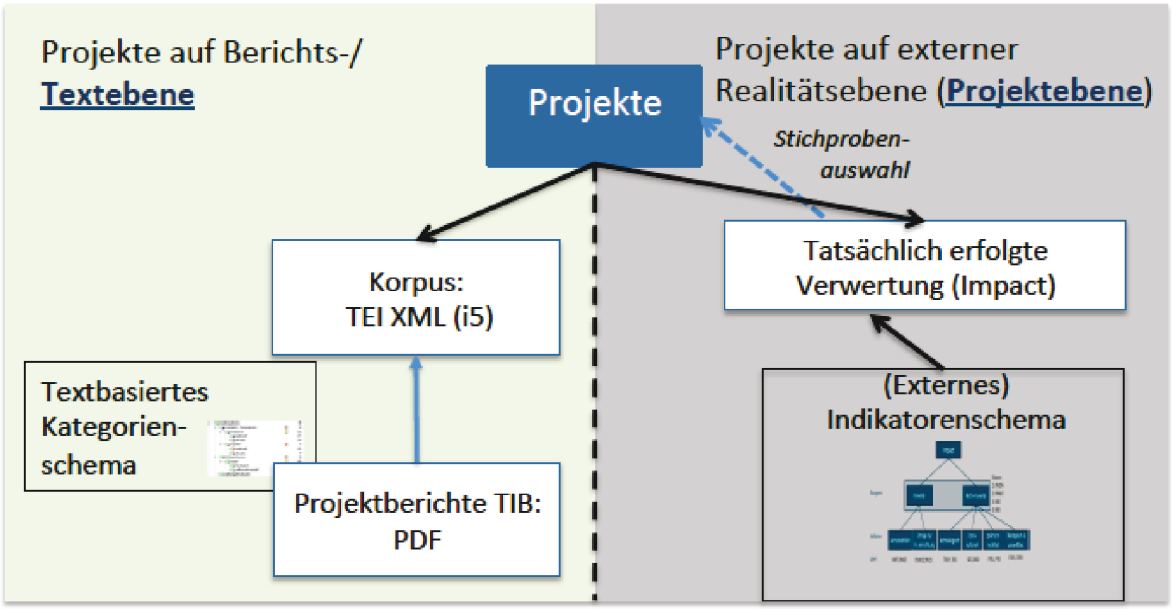

Abb. 2 Projekte mit zwei Betrachtungsebenen: Textebene vs. Projektebene

Einerseits war eine Identifizierung von für oder gegen Transfer sprechenden Indizien auf Seiten der Textebene notwendig, die sich in den (Forschungs-)Projekt(end)berichten manifestiert (vgl. die linke Seite der Abb. 2). Dazu wurden Berichte als solche gelesen und auf Textebene manuell bezüglich ihres Transfer- und Impactpotenzials annotiert. Andererseits finden sich solche Hinweise auf Seiten der Projektebene, die sich mit dem tatsächlichen Transfer befasst haben und die sich während der Laufzeit oder nach Abschluss eines Projektes in Form von realen Effekten bzw. durch tatsächlich erfolgten Transfer in der Realität nachweisen lassen (vgl. die rechte Seite der Abb. 2). Auf dieser Ebene war eine umfassende Recherche notwendig. Weitere Details hierzu finden sich in AP 2.

AP1 wurde plangemäß Ende 2017 abgeschlossen.

3.1.2. AP2: Stichprobe

Der methodische Ansatz des Verbundes sah die Generierung einer geeigneten Stichprobe vor, die als Bezugsbasis für das maschinelle Lernen dienen sollte. Hierfür griffen die Prozesse der physischen

FKZ 01IO1634
Version 1.0

Aufbereitung (Konvertierung und Strukturierung) durch den Projektpartner TIB sowie der informationellen (Auswahl, Annotation) Aufbereitung durch den Projektpartner IDS ineinander.

Mit den in AP 1 vorgenommenen Eingrenzungen lag dem Projekt eine *Grundgesamtheit* an zunächst als pdf-Dokumente archivierten Textdaten unterschiedlicher Qualität bis hin zum reinen Digitalisat vor. Die Annahme zu Beginn der Bearbeitung von AP2 war, dass die Grundgesamtheit selbst die Stichprobe darstellen kann, wenn für alle enthaltenen Projekte eine belastbare Transferaussage getroffen werden kann, die sich auf Hinweise tatsächlich erfolgten Transfers (vgl. Abb. 2 rechte Seite) nach Abschluss des Projektes stützt.

Der tatsächlich erfolgte Transfer auf Projektebene (externes Transfer- und Impactpotenzial) wird dabei anhand des aus Projekten heraus entstandenen Impacts gemessen. Transfer, der demnach nicht mehr einschränkend über bestimmte Formate wie Patente oder Lizenzen, sondern allein vom Nachweis seiner *Wirkung* (vgl. Kap. 2.1.1) her gemessen wird, kann sich indes in sehr unterschiedlichen Formen zeigen und wurde für die Messung entsprechend durch den Projektpartner IDS mit Unterstützung von G&K kategorisiert. Hierfür wurden zwei Kategorien „monetärer Impact“ (also Impact, der sich in Geldfluss messen lässt) und „nicht-monetärer Impact“ (also Impact, der sich nicht in Geldfluss messen lässt) festgelegt, die sich in die vier Klassen „monetärer Impact“ (Klasse 1), „nicht-monetärer Impact“ (Klasse 2), „monetärer und nicht-monetärer Impact“ (Klasse 3) und „kein Impact“ (Klasse 4) ausprägen können. Diese beiden Kategorien lassen sich aus verschiedenen, im Zuge der Recherche auftretenden Indikatoren herleiten, die wiederum jeweils in zwei Ausprägungen (Labels: nachweisbar bzw. nicht nachweisbar) vorliegen können. Der hier beschriebene Zusammenhang ist in der nachfolgenden Abbildung (Abb. 3) dargestellt.

FKZ 01IO1634
Version 1.0

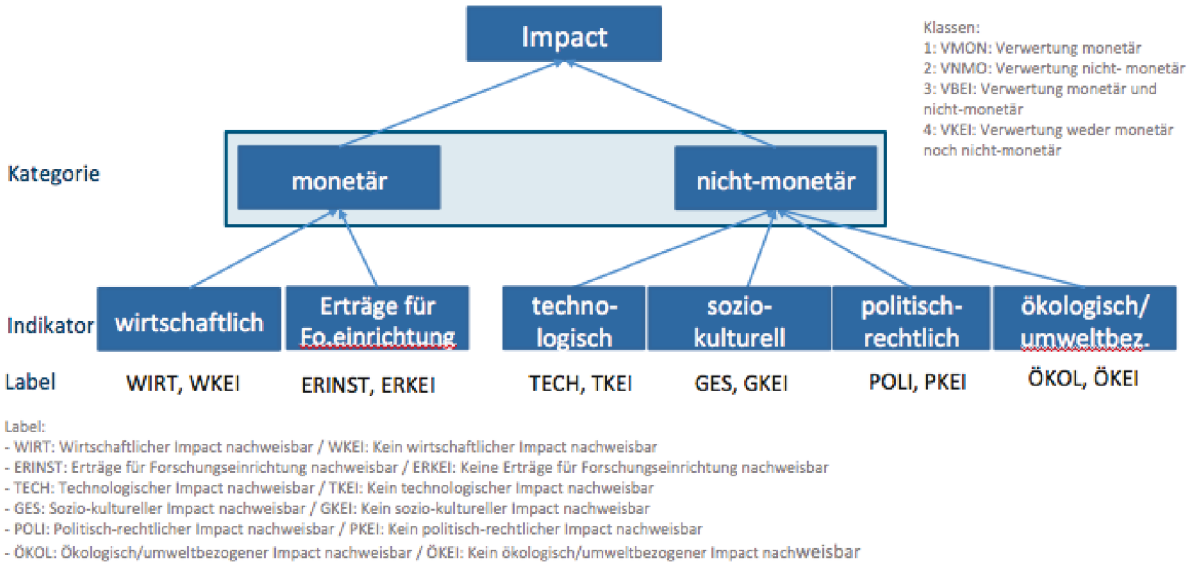

Abb. 3 Kategorisierung des tatsächlich erfolgten Transfers anhand von Impact (Projektebene)

Ein Projekt kann dabei mehrere Formen von Impact aufweisen, den einzelnen Indikatoren darf jedoch nur ein Label zugewiesen werden (z.B. entweder WIRT oder WKEI). Laut Schema werden die Kategorien monetärer Impact bzw. nicht-monetärer Impact nicht separat erhoben, sondern leiten sich aus den sechs erfragten Indikatoren wirtschaftlich, Erträge für Forschungseinrichtung, technologisch, sozio-kulturell, politisch-rechtlich und ökologisch bzw. umweltbezogen ab.

Zu Beginn der im Verantwortungsbereich von IDS und G&K angesiedelten Recherche tatsächlich erfolgten Transfers wurde zunächst davon ausgegangen, dass die Informationsgewinnung zum Nachweis erfolgten Transfers über messbare Impact-Effekte ausgewählter Projekte unter Berücksichtigung von Aufwand und Ergebnis im Rahmen von *TextTransfer (Pilot)* mittels Onlinerecherche zu bewältigen sei.

Wie sich jedoch herausstellte, gestaltete sich dieser Ansatz der Sekundärrecherche der Informationsgewinnung als ein sehr aufwändiger Prozess mit einem hohen Ungenauigkeitsrisiko, u.a.

FKZ 01IO1634
Version 1.0

da Projektnamen mehrfach verwendet werden, der eindeutige Bezug zwischen Projekten und Produkten unklar bleibt etc.

Der Rechercheansatz wurde daher in einer zweiten Phase zugunsten einer Primärerhebung durch telefonische Interviews angepasst.

Um Schritt 2 realisieren zu können, mussten zu allen Projekten der Grundgesamtheit die Leiter und Leiterinnen der Gesamt- oder Teilprojekte identifiziert und deren Emailadressen und Telefonnummern recherchiert werden. Bei Projekten, bei denen diese Informationen bereits Bestandteil der Abschlussberichte waren, gestaltete sich diese Informationsbeschaffung als relativ einfach. Da diese Angaben jedoch keine formale Anforderung bei der Erstellung der Berichte darstellen, erwies sich auch diese Recherche teilweise unter Berücksichtigung von Aufwand und Ergebnis als sehr aufwändig.

Die final identifizierten Projektleiterinnen und Projektleiter bzw. maßgeblichen Projektmitarbeiterinnen und Projektmitarbeiter, wobei jeweils eine Person stellvertretend für das Gesamtprojekt ausgewählt wurde, wurden per Email mit der Bitte um ein ca. 15-minütiges telefonisches Interview kontaktiert – mit explizitem Hinweis auf die Einhaltung der Datenschutzbestimmungen bzgl. der Anonymität der aus der Datenerhebung gewonnenen Erkenntnisse.

Auch wenn dieser Ansatz sich als ebenfalls recht aufwändig erwies (u.a. durch die Unerreichbarkeit der Ansprechpersonen, Änderung der Zuständigkeiten, Jobwechsel etc.), konnten auf diese Art genauere Ergebnisse zu tatsächlich erfolgter Verwertung recherchiert werden bzw. klare, eindeutige Zuordnungen in eine oder beide Impact-Kategorien erfolgen (vgl. hierzu auch *Anlage 4.1 Leitfaden der telefonischen Interviews*).

Um in dem gesetzten Zeitrahmen Ergebnisse für die weitere Bearbeitung zu erhalten, war eine weitere Annahme, dass die Ausprägungen tatsächlich erfolgten Transfers, die sogenannten Labels, eines Projektpartners und dessen Einzelprojektberichts sich auch auf die anderen Berichte des

FKZ 01IO1634
Version 1.0

Projektes übertragen lassen, da es sich immer um Verbundergebnisse handelte, die nicht isoliert betrachtet werden können.

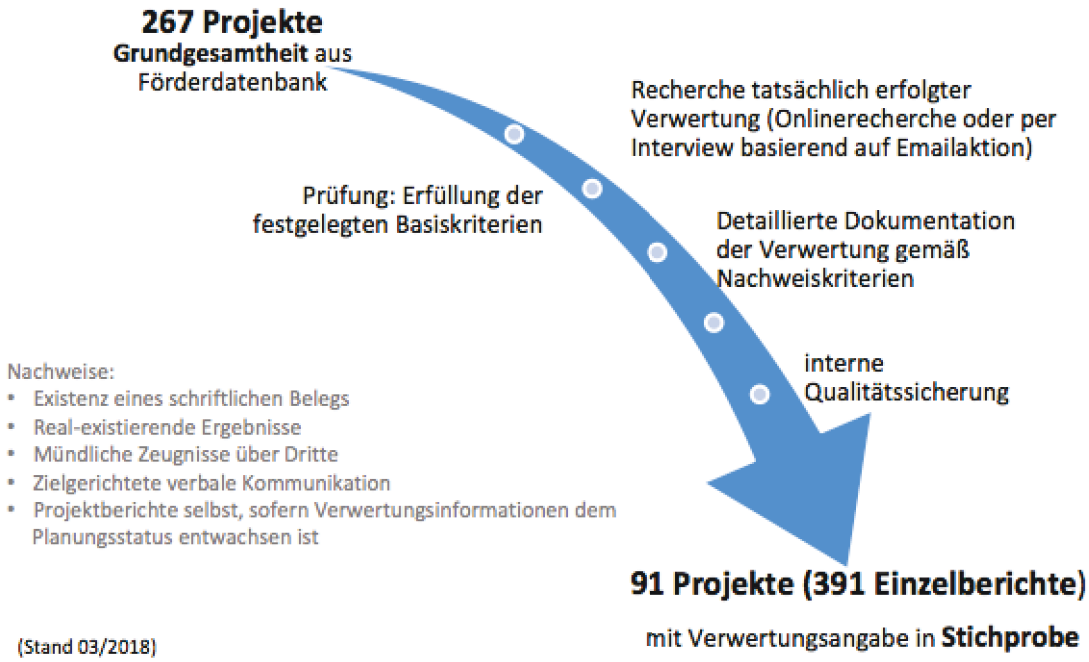

Abb. 4 Stichprobe Mobilität – Projekte mit Nachweis tatsächlich erfolgten Transfers anhand von Impact

Die Informationsgewinnung tatsächlich erfolgten Transfers mittels Interviews wurde plangemäß im ersten Quartal 2018 mit dem Ergebnis der Identifikation folgender Zuweisungen für die insgesamt 91 Projekte mit ihren insgesamt 391 Einzelberichten abgeschlossen, die damit auch die finale Stichprobe zur weiteren Bearbeitung darstellten (vgl. hierzu im Folgenden Kap. 3.1.5):

- **36 Projekten** mit insgesamt 168 Einzelberichten konnte die Kategorie Nicht-monetärer Impact (Klasse 2: **VNMO)** im technologischen, sozio-kulturellen, politisch-rechtlichem und/oder ökologischen bzw. umweltbezogenen Bereich zugewiesen werden.
- **40 Projekten** mit insgesamt 173 Einzelberichten konnte sowohl monetärer Impact im wirtschaftlichen Bereich oder Erträge für die Forschungseinrichtung als auch nicht-monetärer Impact im technologischen, sozio-kulturellen, politisch-rechtlichem

und/oder ökologischen bzw. umweltbezogenen Bereich zugewiesen werden (Klasse 3: **VBEI).**

- **15 Projekten** mit insgesamt 50 Einzelberichten konnte weder monetärer Impact im wirtschaftlichen Bereich oder Erträge für die Forschungseinrichtung noch nicht-monetärer Impact im technologischen, sozio-kulturellen, politisch-rechtlichem und/oder ökologischen bzw. umweltbezogenen Bereich zugewiesen werden (Klasse 4: **VKEI).**
- Ausschließlich monetärer Impact im wirtschaftlichen Bereich oder Erträge für die Forschungseinrichtung konnte keinem Projekt zugewiesen werden (Klasse 1: **VMON**).

Da es sich bei der Quellenbasis der TIB um PDF-Dokumente handelt, mussten in einem nächsten Schritt die Einzelberichte für die weitere Bearbeitung durch den Projektpartner TIB in das maschinenlesbare Format TEI XML (i5), basierend auf einer eigens hierfür konzipierten Arbeitsanleitung (vgl. hierzu auch *Anlage 4.2 Arbeitsanleitung für Konvertierungsarbeiten TIB*), konvertiert werden. Eine besondere Herausforderung war hierbei, dass die Dateien sich nicht vollständig automatisiert konvertieren ließen, sondern immer manueller und intellektueller Aufwand pro Einzelbericht nötig war. Zudem barg die Extraktion von annotierbarem Text in TEI XML (i5), dem im Rahmen von *TextTransfer (Pilot)* gefordertem Format zur Weiterverarbeitung, die Gefahr des Informationsverlustes. Eine manuelle und intellektuelle Nachbearbeitung, die je nach Bericht abhängig von Umfang, Formatierung und Struktur von zwei Stunden bis zu mehreren Arbeitstagen (beispielsweise u.a. auf Grund von Abschlussberichten in Form von Bilddateien) betragen konnte, war daher unumgänglich.

Eine Untersuchung automatisierter Konvertierungsprozesse an der TIB ergab, dass eine weitgehende Automatisierung der Konvertierung Einbußen bezüglich der Qualität des TEI-Markup bedeuten würde, die nicht akzeptabel wären.

Sobald die Daten der Stichprobe in das offizielle XML-Format konvertiert waren, wurden sie in einem nächsten Schritt vom IDS in ein erweitertes XML-Format aufbereitet. Das neue XML-Format beinhaltete, wie oben bereits beschrieben, sowohl die Informationen zu externen Impactkategorien auf Projektebene als auch die Annotation mit den textbasierten Impactkategorien (relevante Textinstanzen gemeinsam mit deren Haupt- und Unterkategorien, vgl. hierzu untenstehende Details).

FKZ 01IO1634
Version 1.0

Da sich die Konvertierung der Stichprobe in das offizielle XML-Format als deutlich aufwändiger erwies als zu Beginn des Projektes angenommen, musste die Stichprobe durch das IDS so aufbereitet werden, dass aus den PDFs der Textinhalt mit Standard-Tools extrahiert und dieser wiederum in einem weiteren, aufwändigen manuellen Schritt bereinigt und formatiert werden konnte. Nach dieser Konvertierung lagen die meisten Berichte als Text-Dateien für weitere Verarbeitungsschritte vor – trotz einiger Qualitäts- und Inhaltsverluste beim Konvertierungsprozess.

Parallel zu der Analyse von Transferinformationen auf Projektebene (vgl. hierzu auch im Folgenden unter AP 5 Deduktiver, Top-Down-Ansatz: Von der Theorie zu den Daten) wurde die Stichprobe zusätzlich mit einschlägigen Impactinformationen ab dem dritten Quartal 2018 durch das IDS auf der Textebene (vgl. hierzu auch im Folgenden unter AP 5: Induktiver, Bottom-Up-Ansatz: Von den Daten zur Theorie) analysiert. Ausgangspunkt dieses Annotationsansatzes war es, dass eine Reihe von Berichten, die nach dem Zufallsprinzip aus der Stichprobe ausgewählt wurden, in einem ersten Schritt aus rein subjektiver Perspektive von insgesamt drei menschlichen Bearbeitern gründlich gelesen und annotiert wurden, um die Textstellen (Sätze oder Abschnitte) zu markieren, die aus subjektiver Sicht Rückschlüsse auf Transfer- und Impactpotenzial eines Projektes geben. Da die Berichte alle in deutscher Sprache gehalten sind, wurden die ausgewählten Stichproben außerdem, um verallgemeinerbare Kategorien für Nicht-Muttersprachler zu definieren, ins Englische übersetzt und drei weiteren menschlichen Annotatoren (Englischsprachige) vorgelegt, um das gleiche Verfahren durchzuführen. Den Annotierenden lagen bei diesem Ansatz zu keinem Zeitpunkt die im ersten Ansatz recherchierten tatsächlich nachgewiesenen Transfer- und Impactpotenziale eines Projektes vor, so dass letztendlich eine relativ objektive Annotation auf Grund des textbasierten Informationsgehalts eines Projektes erfolgen konnte. Basierend auf diesen ersten Ergebnissen wurde ein erstes Codebook entwickelt, das in mehreren Phasen (Annotation, Evaluation, Neuanpassung) verfeinert wurde und in der Folge für die Annotation zum Transfer- und Impactpotenzial der Projekte auf Textebene diente.

Für die Annotation auf Textebene wurde pro Projekt der Stichprobe ein Dokument ausgewählt, in dem alle Textteile markiert wurden, die potenziell für eine Identifikation von Transfer- und

FKZ 01IO1634
Version 1.0

Impactpotenzial relevant sein konnten. Priorisiert wurde dafür ein Gesamtbericht verwendet, der die Projektergebnisse aller Partner beinhaltete (vgl. hierzu *Anlage 3.4 Textbasierte Annotation: Auswahl und Charakteristika der Berichte*).

Die relevanten Teile des Dokumentes wurden in einem ersten Schritt jeweils von zwei unabhängigen Annotierenden mit den entsprechenden Haupt- und Unterkategorien versehen. Die in Abbildung 5 dargestellten Haupt- und Unterkategorien standen den Annotierenden dabei zur Verfügung:

- **1. Impact auf Domäne/Bereich/Feld**
  - Impact auf die Wirtschaft (wirt)
  - Impact auf die Umwelt (umwelt)
  - Impact auf die Gesundheit im Allgemeinen (gesund-all)
  - Impact auf Technik & Technologie (tech)
- **2. Impact auf Gesellschaft, öffentliche Meinung und/oder Wertesystem**
  - Impact auf Judikative und Legislative (legis)
  - Impact auf Gesundheitswesen als Einrichtung (gesund-sys)
  - Impact auf öffentliches Bildungswesen, (Aus)Bildung/Erziehung (bildung)
  - Impact auf Berufswelt (beruf)
  - Impact auf politische/soziale Themen (pol-sozial)
  - Impact auf Bewusstsein/Wahrnehmung (bewusst)
- **3. Impact Outcome**
  - Impact in Form von realen Produkten bzw. deren Prototypen (produkt)
  - Wissensbasierter Impact (wissen)
  - Impact in Form von Richtlinien/Guidelines (richt)
  - Sonstiger Impact (sonst)
- **4. Eigenschaften/Features von Impact Outcome**
  - Neuheit (neu)
  - Sicherheit (sicher)
  - (Daten)Schutz (datenschutz)
  - Nachhaltigkeit (nachhaltig)
  - Flexibilität (flex)
  - Personalisierung (person)
  - Sonstiger Impact (sonst)

Abb. 5 Codebook-Kategorien für die textbasierte Annotation

Zusätzlich zu den in Abb. 5 aufgelisteten detaillierten Kategorien hatten die Annotierenden die Möglichkeit, die Optionen „Viele“, „Sonstige“ und „Kein Impact“ zu vergeben.

FKZ 01IO1634
Version 1.0

Folgende Vorgaben waren bei der textbasierten Annotation insbesondere zu beachten:

- Die Annotation fand auf Satzebene statt.
- Ziel war es, eine eindeutige Aussage bzgl. einer der Impact-Hauptkategorien Domäne, Gesellschaft, Outcome oder Feature zuzüglich einer der pro Hauptkategorie zugehörigen möglichen Unterkategorie pro Satz zu erhalten.
- Für jeden Impact-relevanten Satz sollte die am besten geeignete Haupt- und Unterkategorie ausgewählt werden, auch wenn jeweils mehrere Kategorien in Frage kamen.
- Wenn ein Satz zwar eine mögliche Impact-Hauptkategorie enthielt, diese jedoch nicht Teil der vorgegebenen Hauptkategorien war, sollte dieses Merkmal mit "Sonstige" vermerkt werden.
- Einem Satz, der keinen Impact enthielt, sollte die Hauptkategorie „Kein Impact“ zugewiesen werden und keine Unterkategorie.
- Ein Satz, für den zwei oder gar mehrere Kategorien gleichermaßen geeignet schienen, wurde mit der Kategorie "Viele" versehen.
- Eine Absprache während einer laufenden Annotation durch die Annotierenden war explizit untersagt.

Das ausführliche Codebook findet sich in der Anlage unter Punkt 4.3.

Nach Abschluss der Annotation durch die insgesamt vier Annotierenden, die in sechs Paarkombinationen arbeiteten, gab es bei den insgesamt 6.384 zu annotierenden Sätzen 60% Übereinstimmung. 2.576 Sätze bzw. 40% wiesen entweder keine Übereinstimmung bei den Annotationen auf oder waren nicht eindeutig annotiert.

Der Anteil der Sätze ohne eindeutige Kategorisierung („Viele“, insg. 40% aller Sätze) oder unterschiedlichen Haupt- und/oder Unterkategorien mussten daher adjudiziert werden[7]. Die Adjudikation wurde von zwei weiteren Personen durchgeführt, die nicht an der ursprünglichen Annotation teilgenommen hatten. Nach der Adjudikation lag eine eindeutige Annotation vor, die als Basis für das maschinelle Lernen dienen konnte (vgl. AP 5):

[7] Die Adjudikation dient der Auflösung von Annotationsunterschieden, mit dem Ziel, eindeutige Ergebnisse für das maschinelle Lernen zu erhalten.

Hauptkategorie	Unterkategorie		
Domäne:		**1.016**	
	- gesund-all: 14		
	- tech: 387		
	- umwelt: 47		
	- wirt: 568		
Gesellschaft		**244**	
	- beruf: 14		
	- bewusst: 145		
	- bildung: 50		
	- legis: 5		
	- pol-sozial: 30		
	- gesund-sys: 0		
Outcome:		**984**	
	- produkt: 407		
	- richt: 23		
	- wissen: 554		
Feature:		**182**	
	- datenschutz: 5		
neu	- flex: 24		
	- nachhaltig: 21		
	- neu: 76		
neu	- person: 12		
	- sicher: 44		
Summe der Sätze mit Impact Gesamt:		**2.426**	**38%**
Summe Sätze ohne Impact:		**3.675**	**58%**
Summe Sätze "offene Fälle"		**283**	**4%**
Gesamtsumme der zu annotierenden Sätze		**6.384**	

Abb. 6 Ergebnis der textbasierten Annotation

Bis auf „gesund-sys“ hatten sich alle Unterkategorien als relevant erwiesen. Die Unterkategorien „flex“ und „person“ aus der Hauptkategorie „Feature“ waren ursprünglich nicht Teil des Codebooks, sondern wurden nachträglich im Rahmen der Adjudikation auf Grund des augenfällig relativ häufigen Vorkommens beider Unterkategorien hinzugefügt.

Dank der Annotation und der anschließenden Adjudikation konnte in 38% der Fälle (2.426 Sätze) eine eindeutige Impactzuweisung auf eine der Hauptkategorien Domäne, Gesellschaft, Outcome oder Feature mit einer entsprechenden Unterkategorie zugewiesen werden. 58% der Fälle (3.675 Sätze) wurden eindeutig mit der Kategorie „Kein Impact“ versehen. 4% der Fälle (283 Sätze) wurden übereinstimmend viele Haupt- oder Unterkategorie zugewiesen. Diese wurden mit „offen“

FKZ 01IO1634
Version 1.0

gekennzeichnet und konnten auf Grund ihrer Nichteindeutigkeit für das maschinelle Lernverfahren nicht berücksichtigt werden.

Das Codebook in seiner vorliegenden Form erwies sich insgesamt als Basis für die textbasierte Annotation als gut geeignet, trotz der kleineren Modifikationen, die sich auf Grund der Ergebnisse der individuellen unabhängigen Annotationen ergeben haben.

Wie bei Annotationsaufgaben üblich wurden Kappa-Werte, die die Differenz zwischen tatsächlicher und zufälliger Übereinstimmung normieren, für die einzelnen Annotierenden-Paare berechnet, um einen objektiven Eindruck hinsichtlich der Zuverlässigkeit der Ergebnisse der Annotation zu erhalten. Die Kappa-Werte bezogen sich auf die genannten Hauptkategorien Domäne, Gesellschaft, Outcome, Feature, Kein Impact, Viele und Sonstige.

Annotierende-Paare	Übereinstimmung	Kappa-Wert
anno 01 anno 02	0,89	**0,79**
anno 01 anno 03	0,74	**0,61**
anno 01 anno 04	0,51	0,32
anno 02 anno 03	0,80	**0,69**
anno 02 anno 04	0,46	0,29
anno 03 anno 04	0,43	0,23

Abb. 7 Kappa-Werte der Annotierenden-Paare der textbasierten Annotation

Wie in Abb. 7 ersichtlich erzielten drei Annotierende (anno 01, anno 02, anno 03) gute Übereinstimmungen und dementsprechend gute Kappa-Werte (0,6 und höher) untereinander. Die Paare, die mit anno 04 gebildet wurden, hatten dagegen moderate Kappa-Werte (unter 0,6). Eine Erklärung wird in einer wesentlich kürzeren Einarbeitungszeit von anno 04 gesehen und den damit verbundenen geringeren Erfahrungswerten. Ein weiterer Faktor, der eine wichtige Rolle bei der Berechnung der Kappa-Werte spielt, stellt das Verhältnis der Kategorien untereinander dar. In dem

FKZ 01IO1634
Version 1.0

vorliegenden Fall war deren Verteilung sehr unterschiedlich (vgl. hierzu auch Abb. 6). Der Anteil der Sätze, die mit „Kein Impact“ annotiert wurden, betrug mehr als die Hälfte aller Sätze. Weiterhin waren die Kategorien „Domäne“ und „Outcome“ viel frequenter als die Kategorien „Gesellschaft“ und „Feature“. Das ungleiche Verhältnis der Kategorienhäufigkeit leistete somit einen Beitrag zu den Kappa-Werten: Der durchschnittlicher Kappa-Wert über alle Paare verteilt beträgt 0.48 und kann als „moderat“ bezeichnet werden.

Mit Blick auf das maschinelle Lernen wurde der Ansatz der textbasierten Annotation mittels der Ergebnisse aus dem deduktiven, in der Realität nachgewiesenen Transfer- und Impactpotenzial auf seine Aussagekraft hin überprüft (vgl. hierzu u.a. AP5 und AP6).

Wie zu Beginn dieses Kapitels beschrieben gingen mit der Stichprobengenerierung, die für das maschinelle Lernen Grundvoraussetzung war, erhebliche Herausforderungen der Konvertierung der Rohdatenbasis einher. In der Folge hat der Projektpartner TIB Anpassungen im Workflow der Konvertierung in das vom IDS geforderte Format TEI XML (i5) vorgenommen und mit Start des Arbeitspaketes antragsgemäße personelle Kapazitäten aufgebaut. Wesentliche Herausforderungen bei der Durchführung des Arbeitspaketes ergaben sich aus nicht standardisierten Abgabeformaten eingereichter Berichtstexte - eine Hürde für das maschinelle Lernen, die in einem potenziellen Nachfolgeprojekt ein wichtiger Arbeitsschwerpunkt sein wird. 23 der zu in TEI XML (i5) konvertierenden Berichte (knapp 6%) waren durch unverhältnismäßig hohen Arbeitsaufwand zu Projektende noch offen. Um Aufwand und Ergebnis in Relation zu halten wurde im Projektteam beschlossen, die Restkonvertierung in einer zweiten Projektphase vorzunehmen. Die Experimente im Rahmen des maschinellen Lernens (vgl. hierzu Kapitel *3.1.5 AP5: Softwareanpassung*) blieben davon unberührt und konnten wie geplant – unter Hinzuziehung der durch das IDS mit Standard-Tools erstellten Text-Dateien (vgl. hierzu ebenfalls die Ausführungen in diesem Kapitel weiter oben) – durchgeführt werden, trotz einiger Qualitäts- und Inhaltsverluste der auf diesem Wege konvertierten Dateien.

AP2 war bei Projektende gemäß Anpassungen im Projektplan (vgl. Kap. 3.1.3 *AP 3: Inventar*) abgeschlossen.

FKZ 01IO1634
Version 1.0

3.1.3. AP3: Inventar

Ursprünglich sah der Projektantrag ein Arbeitspaket vor, das ein deutschsprachiges Inventar an Schlüsselbegriffen und Kriterien entwickeln sollte und das Voraussetzung für eine semantische Annotation der gewählten Stichprobe sein sollte. Das Inventar sollte dabei so zu gestalten sein, dass das Spektrum an Begrifflichkeiten im Bereich des Wissenstransfers in seiner gesamten Heterogenität berücksichtigt und der Varianz gewählter Themenfelder gerecht werden würde.

In der Praxis zeigte sich nach ersten Tests dieses Ansatzes im Laufe des ersten Berichtszeitraumes sehr schnell, dass maschinelle Lernverfahren tatsächlich keine derart feingranulare und kontextsensible Annotationsvorgaben benötigen und eigenständige automatisierte Durchläufe sehr viel flexibler Texteigenschaften und Schlüsselbegriffe abgleichen können. Die Entwicklung von Kategorienschemata zum Transfer- und Impactpotenzial auf externer Projektebene als auch auf textbasierter Ebene, wie sie in AP 2 dargestellt wurden, erwies sich hingegen als effizienterer Weg im Umgang mit der Stichprobe als ein in der Erstellung einerseits zu umfangreich und in der Skalierbarkeit als zu sperrig anzusprechendes Inventar.

3.1.4. AP4: Anwendungsfälle

Es oblag *TextTransfer (Pilot)* zunächst, anhand exemplarischer eingegrenzter Laborbedingungen ein grundsätzliches Funktionieren des hier zu erprobenden Ansatzes nachzuweisen. Die Eingrenzung erfolgte hierbei domänenspezifisch über eine kriteriengestützte Definition der Quellenauswahl seitens des IDS und quelltypspezifische, durch die Identifikation technisch wie physisch für das maschinelle Lernen geeigneter, Textformate. Aufgabe des Arbeitspaketes war somit die Entwicklung von exemplarischen thematischen Anwendungsfällen. Die Projektpartner erzeugten unter den genannten Rahmenbedingungen (vgl. AP 1) eine Stichprobe aus der Domäne Mobilität, die für das Projekt als Anwendungsfall fungierte. Perspektivisch ergeben sich weitere Einsatzpotenziale für die Methode aus domänen- wie quelltypspezifischen Erweiterungen (vgl. AP 7).

Zur Systematisierung möglicher Anwendungsfälle für *TextTransfer (Pilot)* wurde im Rahmen des Projekts das im Folgenden dargestellte morphologische Tableau entwickelt:

FKZ 01IO1634
Version 1.0

Merkmale	Ausprägungen				
Anwender	Transfer-beauftragte	Direktoren	Wissenschaftler	...	...
Tätigkeitsfeld Institut	Mobilität	Künstliche Intelligenz	Geo-Ökonomie	...	...
Disziplinen-schwerpunkt	MINT	GSW	Kunst	...	...
Textarten	Endberichte geförderter Projekte (Bund)	Anträge	Berichte anderer geförderter Projekte	Beiträge in wissen-schaftlichen Journals	...

Abb.8 TextTransfer – Morphologisches Tableau Anwendungsfälle

Als „Merkmale“ werden hier Aspekte bezeichnet, die sich im Projektverlauf als geeignet gezeigt haben, den Einsatzbereich von *TextTransfer* näher zu beschreiben:

- „Anwender“ bezeichnen Personen (bzw. Rollen) in Instituten, die *TextTransfer* selber in einer Organisation einsetzen, mindestens aber die Informationen nutzen können. An dieser Stelle ist anzumerken, dass der eigentliche Einsatz der Methode *TextTransfer* im jetzigen Entwicklungsstadium von *TextTransfer (Pilot)* linguistische Expertise erfordert, die nicht in allen Instituten vorhanden ist.
- Mit Bezug auf das im Rahmen des Sondierungsprojekts gewählte Vorgehen ist zunächst auch das Tätigkeitsfeld eines als potenzieller Anwender agierenden Instituts von Relevanz. So konzentrierten sich die Arbeiten aus den oben genannten Gründen (vgl. AP 1) auf die Domäne „Mobilität“, für die sich als Ergebnis des vorliegenden Pilotprojektes belastbare Aussagen formulieren lassen. Für Forschungsergebnisse aus Instituten, deren Schwerpunkt in anderen Domänen liegen, sind zur Zeit noch keine Prognoseangaben möglich. An dieser Stelle ergibt sich Anknüpfungspotenzial für eine domänenübergreifende Erweiterung der Methode.
- Die untersuchten Quelltypen stellen einen wesentlichen Einflussfaktor für die Anwendungsfälle dar. So wurden im Projekt Endberichte öffentlich (häufig BMBF-) geförderter Projekte untersucht. Diese Berichte folgen in den letzten Jahren einer vorgegebenen Struktur, die Formulierungen gehen in der Regel mit der Intention der Verfasser einher, ein Projekt formal einwandfrei abzuschließen. Insofern ist davon auszugehen, dass sich diese Berichte von den nicht-öffentlichen Berichten, aber auch

FKZ 01IO1634
Version 1.0

von Berichten für andere Förderer, Beiträgen in wissenschaftlichen Journalen, aber auch von Anträgen unterscheiden.

Um die prinzipielle Funktionsfähigkeit der Methode *TextTransfer* nachzuweisen, hat sich das Projektteam entschlossen, die Arbeiten auf die erste Spalte des morphologischen Tableaus (vgl. Abb 8) zu konzentrieren. Daher wurden die Perspektive von Technologietransfer-Beauftragten eingenommen, die sich auf Endberichte von durch den Bund geförderten Projekten mit MINT-Schwerpunkt aus der Domäne „Mobilität“ konzentrieren. Die gewählte Merkmalskombination ist durch eine rote Linie in der Abbildung oben skizziert; andere Kombinationen (z.B. Fokussierung auf die Transferbeauftragten-Perspektive bei anderer Domäne) sind möglich.

Für diesen Anwendungsfall konnte die grundsätzliche Funktionsfähigkeit der Methode *TextTransfer* gezeigt werden. Gleichzeitig ergeben sich aber auch Ansatzpunkte für eine Vertiefung der Methode. So sind insbesondere weitere Domänen, etwa aus dem Bereich der im Rahmen der Hightech-Strategie der Bundesregierung als besonders gesellschaftsrelevant identifizierten Handlungsfelder (Künstliche Intelligenz, Gesundheitswesen, Klima, Migration etc.), mögliche Gegenstände der Erweiterung; es ist aber auch sinnvoll, neben den Projektendberichten weitere Quelltypen auszuwählen und in die Entwicklung einzubeziehen. Für die Evaluation neuer Domänen und charakteristischer Quelltypen wären die in der Pilotphase von *TextTransfer* entwickelten und bewährten Kriterien anzulegen, die das Ergebnis der maschinellen Auswertung als Indikatoren potenzieller Impactrelevanz (Anwendungsforschung, Fokus Technologie, Schlüsselthemen Hightech Strategie, außerakademische Zielgruppen, Transdisziplinarität etc.) besonders zu stimulieren versprechen. Auf diese Weise besteht die Chance, *TextTransfer* zu einem generalisierbaren, textbasierten Prognosewerkzeug des Wissens- und Technologietransfers einzusetzen.

AP4 war zu Projektende planmäßig umgesetzt.

3.1.5. AP5: Softwareanpassung

Zum Zwecke der Softwareanpassung wurde die von den Verbundpartnern kooperativ generierte Stichprobe in aufbereiteter, annotierter und anonymisierter Form an den Projektpartner UIUC zur

FKZ 01IO1634
Version 1.0

Analyse geliefert. Für die Analyse der gewählten Stichprobe war die Evaluation und Anpassung geeigneter Analysewerkzeuge notwendig.

Der Partner UIUC erhielt vom IDS Testdaten aus der Stichprobe, die im Format bereits dem der gesamten final ausgelieferten Stichprobe entsprach (vgl. hierzu Kap. *3.1.2 AP 2: Stichprobe*). Die UIUC benutzte die Testdaten, um das technische Setup für das maschinelle Lernen aufzubauen. Die ersten Schritte beinhalteten die Datenaufbereitung – Erzeugung der Datasets (Spreadsheets) – für das Trainieren von Grundmodellen und der Datenanalyse, um weitere Features für das Training erweiterter Modelle zu identifizieren.

Der Prozess des maschinellen Lernens sah dabei vor, dass

- die Datasets in Training- und Evaluationssets getrennt wurden;
- die Trainingssets benutzt wurden, um unterschiedliche Klassifikatoren zu trainieren;
- die Evaluationssets benutzt wurden, um die Klassifikatoren zu evaluieren;
- die Klassifikatoren mit der besten Performance eingesetzt wurden, um neue Berichte/Projekte in Bezug auf Impact- und Transferkategorien zu klassifizieren.

Zu Beginn des Projektes wurde dabei von einem einzigen Ansatz beim maschinellen Lernen zur Software-Anpassung ausgegangen. Aufgrund der Ergebnisse in AP2, das eine zweigeteilte Betrachtung des Impact- und Transferpotenzials über eine projektbasierte bzw. textbasierte Ebene vornahm, musste sich dieser Ansatz auch in entsprechend getrennten Szenarien der Experimente des maschinellen Lernens widerspiegeln:

> *Stufe 1: Deduktiver, Top-Down Ansatz – Von der Theorie zu den Daten: Maschinelles Lernen mit externen Impact- und Transfer-Kategorien (projektbasiert)*
>
> *Stufe 2 Induktiver, Bottom-Up Ansatz – Von den Daten zur Theorie: Maschinelles Lernen mit textbasierten Impact- und Transfer-Kategorien (textbasiert)*

3.1.5.1. Auswahl und Klassifizierung von Merkmalen

Nachdem alle Datensätze, sowohl die für den deduktiven als auch den induktiven Ansatz, mit Labels versehen waren, wurden beaufsichtigte (supervised) Vorhersagemodelle zur Klassifizierung von

FKZ 01IO1634
Version 1.0

Impactkategorien entwickelt. Aufgrund des kleinen Umfangs der Datensätze (vgl. zum Thema u.a. *Kap 3.1.2 AP2: Stichprobe*) fiel die Entscheidung auf die Verwendung klassischer, merkmalsbasierter Klassifikatoren, nämlich der Support Vector Machine (SVM), dem Gaussian Naive Bayes und dem Random Forest Algorithmus, wobei der Schwerpunkt auf Grund der höchsten Genauigkeit der Modelle auf das SVM-Modell gelegt wurde.

Um Merkmale zu extrahieren, wurden zunächst die Datensätze vorverarbeitet: Zahlen, Symbole, z.B.: Umlaute und Stopp-(Funktions-)Wörter und Wörter, die in weniger als 5 und mehr als 95% der Daten in Erscheinung traten, wurden entfernt.

Zur Erstellung der Klassifikatoren für beide Ansätze wurden folgende drei Sätze von Merkmalen genutzt: (1) lexikalische Merkmale, (2) syntaktische Merkmale (Part of Speech (POS)) und (3) domänenspezifische Merkmale (Impact-Unterkategorien). Darüber hinaus wurden für die lexikalischen Merkmale TF*IDF-Scores der Wörter verwendet (Gleichung (1)):

$$tf - idf\,(t, d) = f_{(t,d)} \times \log 1 + \frac{n}{\mathrm{df}(t)}. \qquad (1)$$

wobei f(t,d) die Häufigkeit des Begriffs t im Dokument d, n die Gesamtzahl der Dokumente im Korpus und df (t) die Anzahl der Dokumente in der Dokumentmenge ist, die den Begriff t enthalten.

Um Unigramme, Bigramme und Trigramme zu extrahieren, wurden tf idf Vectorizer in der Python SKLearn-Bibliothek verwendet. Die Unigramm-Merkmale wurden sowohl beim deduktiven als auch induktivem Ansatz als Basis verwendet; die restlichen Merkmale wurden danach hinzugefügt, um die Nützlichkeit der anderen Merkmale bei der Vorhersage von Auswirkungen zu analysieren.

Um die syntaktischen Merkmale zu extrahieren, wurde TextBlob, Deutsches Paket3 (Loria et al., 2013) verwendet, um damit die Wörter in jedem Datensatz mit ihrem jeweiligen POS zu kennzeichnen, bevor die Texte bereinigt wurden. Anschließend wurde die Anzahl der verschiedenen Sätze wie Substantive, Verben, Adjektive, Adverbien usw. in jedem Eintrag gezählt, die dann als

weitere Merkmale zusätzlich zu den lexikalischen Merkmalen hinzugefügt wurden. Für die domänenspezifischen Merkmale wurden die Impact-Unterkategorien als zusätzliche Merkmale in den Vorhersagemodellen verwendet. In den deduktiven Modellen wurden zudem wirtschaftliche, technische, soziokulturelle, politisch-rechtliche oder ökologische bzw. umweltbezogene Unterkategorien hinzugezogen. Für den induktiven Datensatz wurden dagegen Unterkategorien wie Ökonomie, Ökologie, Gesundheit, Technik, Gesetzgebung usw. verwendet (vgl. hierzu auch die Ausführungen in *Kap 3.1.2 AP 2: Stichprobe*). Diese zusätzlichen Merkmale wurde den lexikalischen und syntaktischen Merkmalen hinzugefügt.

Um einer Verzerrung der Instanzen in den Datensätzen entgegenzuwirken, wurde die Synthetic Minority Oversampling TEchnique (SMOTE) verwendet, um damit die Anzahl der Instanzen in kleineren Kategorien zu erhöhen und die Eingabedaten beim Training der Klassifikatoren auszugleichen (Chawla et al., 2002). Im deduktiven Modell ist, wie in Abbildung 10 gezeigt, die Anzahl der Instanzen in der Klasse "Monetäre und nicht-monetäre Auswirkungen" am größten. In den kleineren Klassen, wurden die Instanzen in der Stichprobe übertrieben, um die Eingabedaten auszugleichen. Für den induktiven Modus (Abbildung 12) wurde zunächst die größte Klasse "Kein Impact" unterbewertet, indem 1.000 Sätze nach dem Zufallsprinzip ausgewählt wurden. Danach wurde SMOTE verwendet, um die kleinen Klassen, nämlich "Feature" und "Gesellschaft", synthetisch überzubewerten.

Um die Leistung des Klassifizierers zu erhöhen und eine Redundanz der Merkmale zu vermeiden, wurde schließlich der χ2 (Chi-Quadrat)-Algorithmus (Gleichung (2)) genutzt, um die oberen k (300 < k < 600) Attribute sowohl im deduktiven als auch im induktiven Modell. χ2 wird dabei verwendet, um zu testen, ob das Auftreten eines bestimmten Begriffs und das Auftreten einer bestimmten Klasse unabhängig voneinander sind. Bei einem Dokument D wurde für jeden Begriff die folgende Menge geschätzt und nach ihrer Punktzahl eingeordnet:

$$\square^2(D, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(\square_{e_t e_c} - \square_{e_t e_c})^2}{\square_{e_t e_c}} \qquad (2)$$

FKZ 01IO1634
Version 1.0

In Gleichung (2) zeigt N die beobachtete Frequenz und E die erwartete Frequenz. Wenn das Dokument den Begriff t enthält, nimmt et den Wert 1 an, andernfalls 0. Wenn das Dokument zur Klasse c gehört, nimmt ec den Wert 1 an. Die Werte sowohl für et als auch für ec sind 0, wenn die Regel nicht erfüllt ist.

Zur Implementierung der Algorithmen und Klassifikatoren wurde das Python-Scikit-learn-Paket (Pedregosa et al., 2011) verwendet. Die k-fache Kreuzvalidierung (k = 5) wurde genutzt, um die Modelle zu trainieren, und verwendeten Standardmetriken wie Präzision, Recall, F1, Fläche unter der Betriebskennlinie des Empfängers (ROC AUC), um die Vorhersagegenauigkeit zu bewerten.

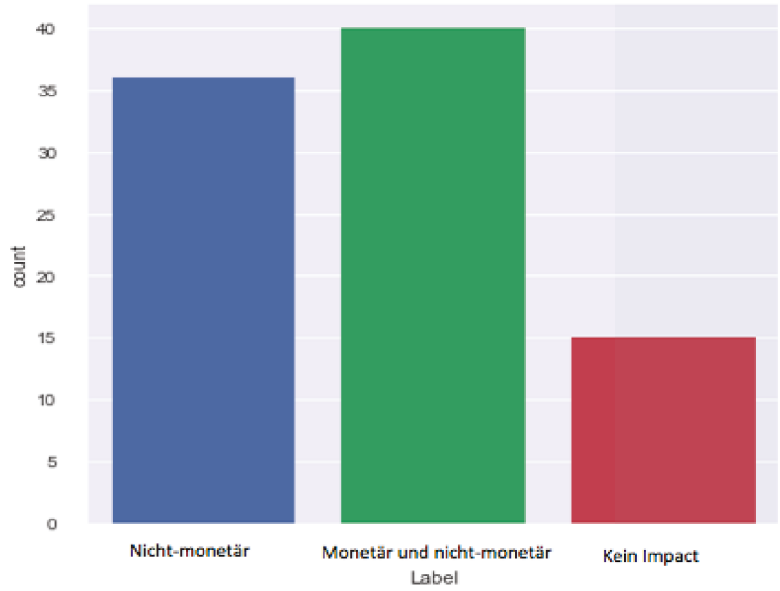

Abb.9 Deduktiver Ansatz – Annotiertes Datenset mit 91 Projekten

Beim deduktiven, überwachten (supervised) Top-Down-Ansatz der Analyse des maschinellen Lernens der Stufe 1 wurden auf Basis des Trainingsset Klassifikatoren trainiert, die anschließend benutzt wurden, um die Projekte aus dem Evaluationsset in Bezug auf externe Impact-Kategorien zu klassifizieren. Insgesamt wurden bei diesem Ansatz 91 Projekte betrachtet (vgl. Abb. 9).

Dabei verteilten sich die Impact-Kategorien und deren Unterkategorien in einem Verhältnis wie in der folgenden Abbildung 10 dargestellt (vgl. hierzu auch *Kap. 3.1.2 AP 2: Stichprobe*):

FKZ 01IO1634
Version 1.0

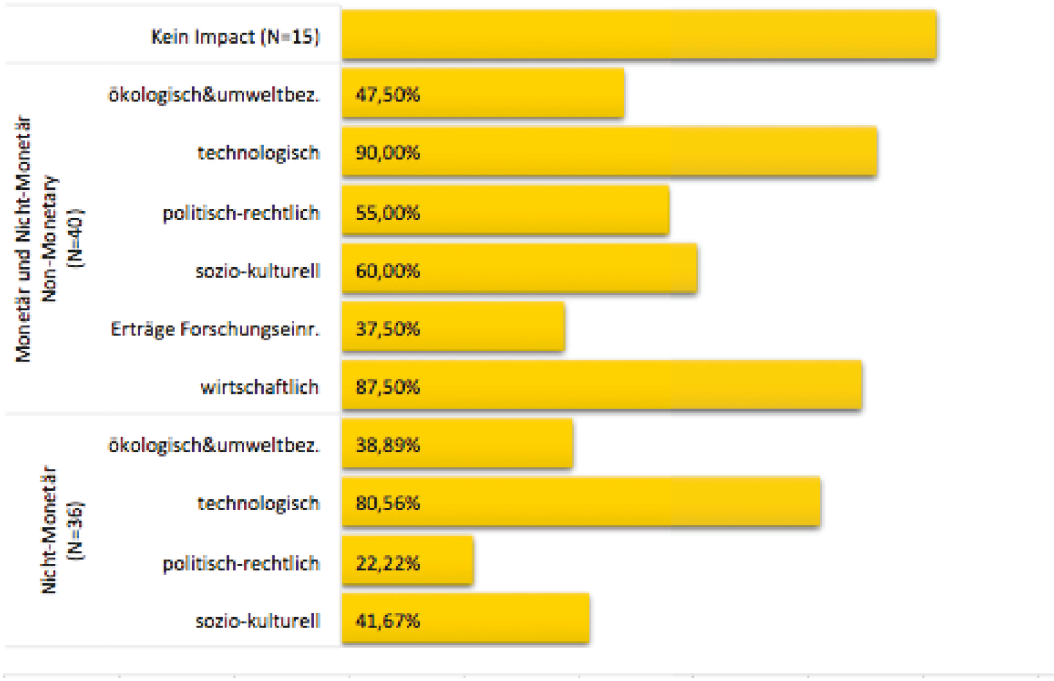

Abb.10 Deduktiver Ansatz – Verteilung der Impact-Kategorien und deren Unterkategorien im annotierten Datenset

Die Basis für den induktiven, Bottom-Up-Ansatz bildete das in AP 2 Stichprobe entwickelte Codebuch, um die Sätze in den einzelnen Berichten der insgesamt 91 Projekte zu kennzeichnen. Die Sätze wurden dabei entsprechend annotiert. Dabei wurden die Textmerkmale, die mit den definierten Kategorien korrelieren, identifiziert. Anschließend wurden überwachte Modelle zur Vorhersage der Impactkategorien auf Satzebene erstellt. (Weitere Details hierzu vgl. (*Kap 3.1.2 AP 2: Stichprobe*). Insgesamt ergab dies ein Datenset mit 2.426 Sätzen für das Maschinelle Lernen (vgl. Abb. 11).

FKZ 01IO1634
Version 1.0

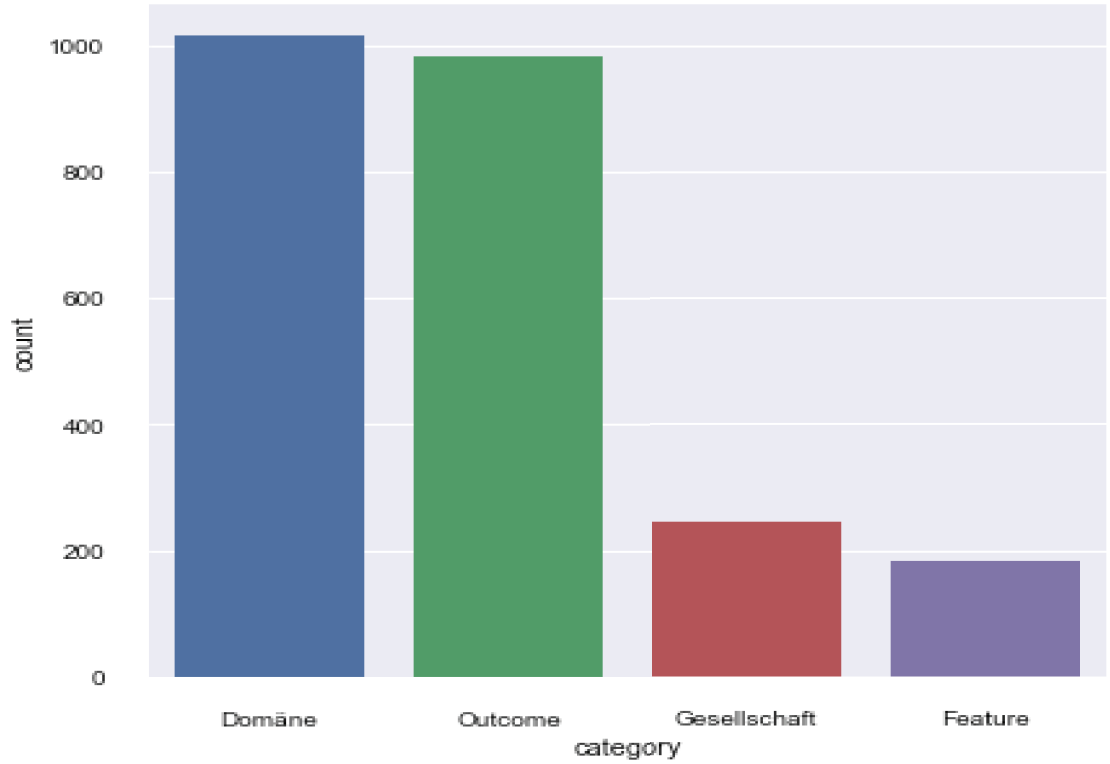

Abb.11 Induktiver Ansatz – Annotiertes Datenset mit 2.426 annotierten Sätzen über 91 Projekte verteilt

Neben der Satzebene wurden auch die annotierten Unterkategorien für das maschinelle Lernen, die wie bereits in diesem Kapitel unter AP 2 beschrieben identifiziert wurden, mit der in der folgenden Abbildung 12 dargestellten Verteilung mitberücksichtigt:

FKZ 01IO1634
Version 1.0

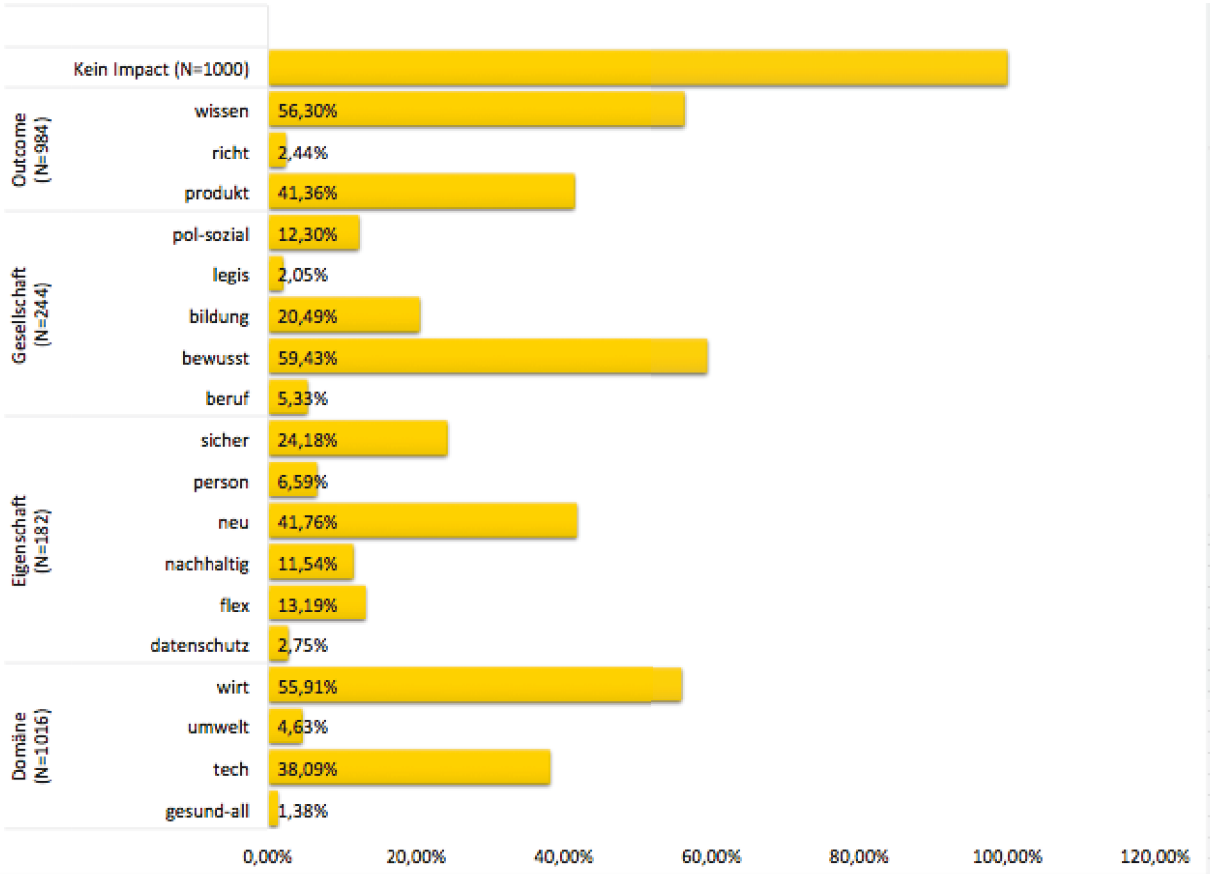

Abb.12 Induktiver Ansatz –Verteilung der Impactkategorien und deren Unterkategorien im annotierten Datenset

Außerdem wurde die Verteilung der induktiven Kategorien auf 91 Projekte, die mit deduktiven Kategorien bezeichnet wurden, analysiert. Wie in Abbildung 13 dargestellt besteht die Mehrheit der Projekte mit "monetärer und nicht-monetärer Impact" aus "Impact auf die Domäne". Die Mehrheit der als "nicht-monetärer Impact" bezeichneten Sätze besteht aus Diskussionen zu Endergebnissen der Forschung wie Produkte, Prototypen, Forschungs- oder Lernmethoden, Richtlinien oder einer anderen Art von Ergebnissen, die als Ergebnis der Forschungsfinanzierung entwickelt wurden. Darüber hinaus wurden "Impact auf Gesellschaft" in Projekten mit "monetärem und nicht-

FKZ 01IO1634
Version 1.0

monetärem Impact" stärker diskutiert. Schließlich besteht ein Projekt ohne monetären oder nicht-monetären Impact ("Kein Impact") aus der geringsten Menge an Diskussionen über Impact auf „Domäne“, „Gesellschaft“, „Outcome“ oder „Feature“. Auch die Korrelation zwischen deduktiven und induktiven Kategorien mit Hilfe der Pearson-Korrelation wurde analysiert, mit dem Ergebnis, dass "Gesellschaft" am meisten mit "Kein Impact", "Domäne" mit "monetärer und nicht-monetärer Impact" (p-Wert < 0,05) korreliert ist.

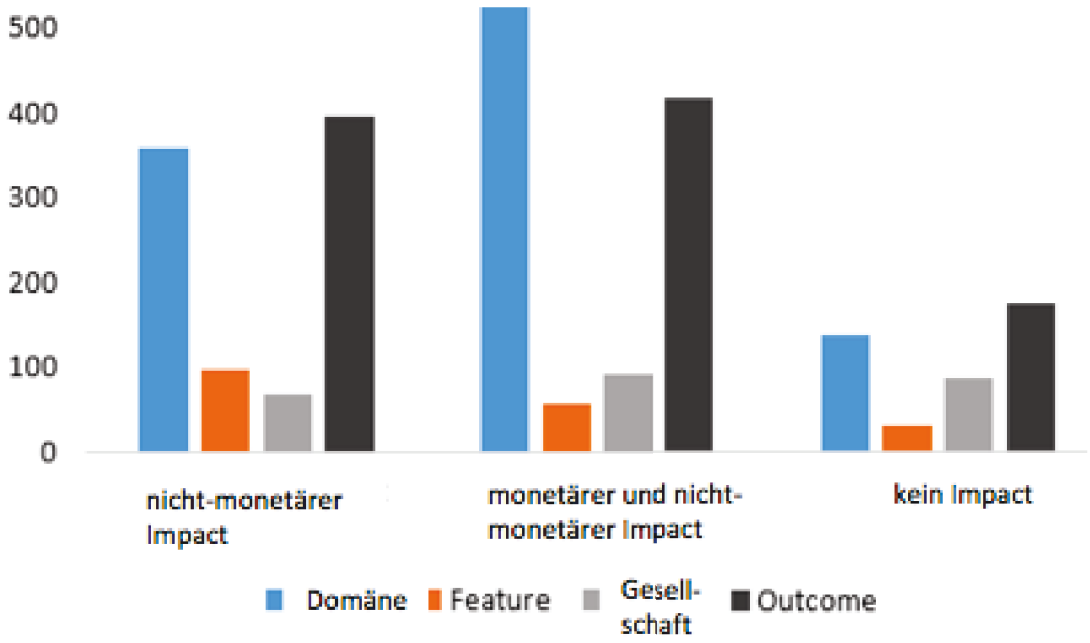

Abb. 13 Verteilung der deduktiven und induktiven Kategorien über die Projekte hinweg

Die Arbeiten im Arbeitspaket wurden planmäßig beendet. Das maschinelle Lernen für beide Ansätze wurde mit einem positiven Ergebnis abgeschlossen. Details hierzu finden sich nachfolgend unter *Kap 3.1.6 AP6: Funktionsnachweis der Methode*.

FKZ 01IO1634
Version 1.0

3.1.6. AP6: Funktionsnachweis der Methode

Ziel des Verbundvorhabens war die Entwicklung einer aus experimentellen Verfahren hergeleiteten, funktionsfähigen Methode der Impactanalyse. Die Projektpartner IDS und UIUC arbeiteten an dieser Stelle zusammen, um die Ergebnisse des maschinellen Lernens zu dokumentieren und den Abgleich des Verbundzieles über geeignete Kriterien zu leisten.

Die Experimente des maschinellen Lernens, die in mehreren Läufen über mehrere Monate sowohl für den deduktiven als auch den induktiven Ansatz bezogen auf die Stichprobe Mobilität durchgeführt wurden, sind mit den Werten der besten Modelle, die für beide Ansätze Precision- und Recall-Werte von über 75% haben und damit im vorliegenden Setup in über 75 von 100 Texten die Verwertungskategorie richtig erkennen würden, ohne dass die 100 Texte vorab komplett etikettiert wären, als positiv zu bewerten:

Model	Deductive Model				Inductive Model			
	P	R	F1	ROC	P	R	F1	ROC
Unigram (Baseline)	72.37	65.81	66.45	73.38	55.62	52.06	52.95	68.91
Ngram (unigram+bigram+trigram)	77.83	75.69	75.32	80.01	56.37	52.77	53.83	69.44
Ngram + POS	77.83	75.69	75.32	80.01	56.2	52.59	53.66	69.31
Ngram + POS +Sub-categories	**80.04**	**76.87**	**76.39**	**80.82**	**79.8**	**78.29**	**78.81**	**85.92**

Abb. 14 Ergebnis des SVM-Classifiers für das deduktive und das induktive Model, Precision (P), Recall (R), F1 Score, Receiver Operating Characteristic Curve (ROC AUC) (Angaben in Prozent)

Wie in Abbildung 14 dargestellt wurde zunächst die Basislinie mit Hilfe von Unigrammen erstellt. Für das deduktive Modell erreichte das Basislinienmodell nach dem Abgleich des Datensatzes und der Auswahl der informativsten Merkmale etwa 66,45% F1-Score. Für das induktive Modell ergab das Basislinienmodell einen F1-Score von 52,95%. Durch das Hinzufügen der Bigramme und Trigramme zur Grundlinie wurde die Leistung in beiden Modellen um etwa 10% im deduktiven Modell und 1% im induktiven Modell erhöht. Wie außerdem aus Abbildung 14 hervorgeht, änderte sich die Precision im N-Gramm-Modell zwar nicht, aber der Recall-Score machte einen enormen Sprung. Dies deutet darauf hin, dass das Hinzufügen von Wörtern zu den Merkmalssätzen bei der Vorhersage echter

FKZ 01IO1634
Version 1.0

positiver Ergebnisse und der Erhöhung der Sensibilität des Klassifikators hilfreich war. Das Hinzufügen von syntaktischen Merkmalen über die lexikalischen Merkmale im deduktiven Modell änderte die Leistung des Klassifikators nicht. Beim induktiven Modell profitierte der Klassifikator nicht von den syntaktischen Merkmalen, da er die Leistung um 0,2% verminderte. Die Kombination der lexikalischen, syntaktischen und domänenspezifischen Merkmale (Unterkategorien) schließlich erhöhte die Leistung der Klassifikatoren in beiden Modellen. Wie außerdem aus Abbildung 14 hervorgeht, profitierte das induktive Modell am meisten von der Kombination aller Merkmale und erreichte etwa 78,81% F1-Wert. Die höchste ROC wurde durch die Kombination aller Merkmale erreicht (80,82% deduktive und 85,92% für die induktiven Modelle).

Diese Ergebnisse bieten bezüglich einer Aussage zum Impact von durch die untersuchte Quellenbasis repräsentierten Forschungen insoweit Möglichkeiten der Generalisierbarkeit, als man für die Domäne Mobilität für neue wissenschaftliche Projektendberichte aus einem bestimmten Sprachraum und Kulturkreis mit denselben Ergebnissen rechnen könnte.

Als weiteres Ergebnis lässt sich festhalten, dass sich beide Ansätze, sowohl der deduktive als auch der induktive, zur Detektion des Impact- und Transferpotenzials eignen. Beide sind als voneinander unabhängige Pipelines programmiert. Die technische Dokumentation hierzu findet sich in der Anlage *4.5 TextTransfer Project-Pipeline*.

3.1.7. AP7: Verwertungskonzept

Der Projektverbund hatte sich zum Ziel gesetzt, die Ergebnisse der Methodenentwicklung so aufzubereiten, dass sie perspektivisch in eine konkrete Anwendung überführt werden können. Die Methode befindet sich mit Projektabschluss in prototypischem Stadium. Der Funktionsnachweis wurde bisher innerhalb einer kontrollierten Umgebung mit hohen Erfolgswerten erbracht. Hierbei waren insbesondere datenseitig Quelltyp, Domäne und Kategorisierungsschema vorgegeben (supervised machine learning). Ein breit angelegter Methodeneinsatz als generisches Transferinstrument wird demnach erst nach Ausweitung dieser Parameter möglich werden. Aufgrund der Tatsache, dass die Kompetenz zur Anwendung der Methode *TextTransfer* zunächst an die Projektpartner gebunden sein wird, lässt sich schon jetzt absehen, dass zu diesem Zweck

FKZ 01IO1634
Version 1.0

geeignete organisatorische wie technische Rahmenbedingungen und Prozesse (Pipeline) implementiert werden müssen.

Um die Anwendung von *TextTransfer* innerhalb einer Forschungseinrichtung zu ermöglichen, bietet sich die Entwicklung eines bedarfsgerechten Rollenmodells an, das die Szenarien Wissensbedarf an Verwertungspotenzial der eignen Forschung (operativ und strategisch) sowie Methodenkompetenz abbildet. So lassen sich mit Bezug auf die Kombination „Text – Wissenstransfer" in einem Institut die folgenden Rollen unterscheiden:

- „Archiv": Im Archiv werden Texte gesammelt und aufbewahrt. Es kann sich um eine institutsinterne Gedächtniseinrichtung (wie etwa eine Bibliothek oder Archive für digitale Forschungsdaten o.ä.) handeln; vielfach werden Texte aber auch „einfach" auf einem Server oder in einer Datenbank archiviert.
- „Wissenschaftlerin/Wissenschaftler": Hierbei handelt es sich um die Person, die die Forschung selber durchgeführt hat. Sie ist in der Regel die um die Sache „wissende Person", nicht selten jedoch mit Zeitverträgen ausgestattet und so möglicherweise nach Projektende nicht mehr verfügbar. Im Kontakt mit potenziellen Nutzern und Anwendern erfüllt diese Rolle primär operative Aufgaben bei der Verwertung eigener Forschung.
- „Technology Transfer Officer (TTO)": Der Technology Transfer Officer steht stellvertretend für die Rolle des für den Transfer Zuständigen; dabei sollen hier pragmatisch Wissens- und Technologietransfer (WTT) gleichgesetzt werden, da es in beiden Fällen darum geht, verwertbares Wissen (auch in Form von Technologien) zu identifizieren. Von größerer Bedeutung ist jedoch, ob in einem Institut eine Person vorhanden ist, die für den Wissenstransfer zuständig ist. Dies ist, so zeigt die Praxis, bei weitem nicht in allen Instituten der Fall. Nicht selten ist die Öffentlichkeitsarbeit für den Transfer „mit-zuständig", was dazu führen kann, dass beispielsweise mit *TextTransfer* verbundene Aufgaben (z.B. die Initiierung einer Suche) aus Zeitgründen nicht vorgenommen werden können.
- „Leitung": Hierbei handelt es sich um die Person, die die übergreifende Rolle in einer Forschungseinrichtung innehat. Darüber hinaus ist es nicht unwahrscheinlich, dass diese Rolle auch den Start eines *TextTransfer*-Prozesses übernimmt bzw. im Sinne der Profilbildung strategisches Interesse an der Früherkennung von Verwertbarkeit eigener Forschung hat. So wird es auch nicht selten die für den Institutserfolg verantwortliche Leitung sein, die die Frage nach verwertbaren Forschungsergebnissen aufwirft.

Zur Beschreibung eines Basisprozesses für *TextTransfer* wird zum einen davon ausgegangen, dass das Erkenntnisinteresse auf der textbasierten Analyse des Verwertungspotenzials von in Projektberichten vorliegenden Forschungsergebnissen liegt. Zum anderen muss die Analyse-Pipeline (vgl. *Kap 3.1.6 AP*

FKZ 01IO1634
Version 1.0

6: Softwareanpassung) im Institut implementiert und die die Methode einsetzenden Mitarbeiter und Mitarbeiterinnen müssen entsprechend geschult sein. An dieser Stelle sei allerdings nochmals darauf hingewiesen, dass für den Moment im Rahmen der Pilotphase *TextTransfer* sowohl Methodeneinsatz als auch Ergebnisauswertung linguistische Kompetenzen und damit Assistenz erfordern. Für ein künftiges Verwertungskonzept zum Einsatz bei Forschungseinrichtungen entlang der bisher etablierten Pipeline ist perspektivisch an ergänzende Komponenten wie eine intuitive Benutzeroberfläche für die Anwendung und/oder ein Unterstützungskonzept seitens des IDS zu denken, welche die Anwendung der prototypischen Methode erleichtern sollen. Als erster Schritt auf diesem Weg hat das IDS zum Abschluss dieser Pilotphase eine technische Methodendokumentation vorgelegt, aus der sich im Rahmen einer zum jetzigen Stand lohnenswerten Methodenerweiterung als weitere naheliegende Maßnahme die praktische Implementierung der Methode bei den Projektpartnern selbst ergäbe.

Das gewählte Rollenmodell wird im Zuge einer Vertiefung in einer möglichen Fortsetzung des Projektes zu verfeinern sein; so ist möglicherweise etwa eine Unterscheidung zwischen „Nutzer der Information“, „Auftraggeber“ und „Durchführer“ zu treffen. Darüber hinaus ist es denkbar, auch die Rolle der Zuwendungsgeber für öffentlich geförderte Projekte einzubeziehen. Diese wird beispielsweise dann relevant, wenn im Interesse einer vereinfachten Auswertung, Projektberichte an ein bestimmtes Dateiformat gebunden werden. Diese Differenzierungen variieren jedoch mit der technologischen Weiterentwicklung der Methode und sollen daher hier nicht weiter ausgeführt werden, sondern wären Thema für ein Hauptprojekt *TextTransfer*.

Vor diesem Hintergrund stellt sich ein idealisierter *TextTransfer*-Basisprozess wie folgt dar:

1. Textidentifikation und -vorbereitung: Die zur Analyse vorgesehenen Texte sind zu identifizieren; darüber hinaus zeigte die bisherige Entwicklung der Methode, dass Texte auch für die automatisierte Auswertung aufzubereiten bzw. in eine maschinenlesbare Form zu konvertieren sind.
2. Sofern der erste Schritt vollständig erfüllt ist, folgt die automatisierte Analyse des Verwertungspotenzials. Wie erwähnt ist hierzu linguistische Expertise notwendig, die sicherlich zur Zeit nicht in allen Instituten vorliegt.

FKZ 01IO1634
Version 1.0

3. Insbesondere die Auswertung der Analyse kommt nicht ohne linguistische Kenntnisse aus. Die Interpretation der Ergebnisse erfordert die Übersetzung der Auswertungsdaten in Verwertungspotenzial.
4. Die Ableitung des Verwertungspotenzials orientiert sich zur Zeit an den weiter oben beschriebenen Impact-Kategorien (vgl. Kap. *3.1.2 AP2: Stichprobe*).
5. Idealerweise wird auf Basis des abgeleiteten Verwertungspotenzials ein Vorschlag zum Verwertungsvorgehen abgeleitet. Die Entwicklung dieses Prozessschrittes muss aber ebenfalls der ggf. anstehenden Methodenerweiterung vorbehalten bleiben.

Dieser, auf den ersten Blick einfach wirkende Basisprozess für *TextTransfer* stellt auf der einen Seite die Möglichkeiten des *TextTransfer*-Einsatzes dar, auf der anderen Seite weist er aber auch auf Anforderungen hin. So wird es sicherlich sinnvoll werden, *TextTransfer*-Expertise im Institut fest zu verankern; diese Aufgabe kann – durch entsprechende Weiterbildung – durch einen institutsinternen TTO übernommen werden, aber: diese Rolle muss es im Institut auch geben. Es wird im Zuge einer potenziellen Methodenerweiterung zu prüfen sein, wie unterschiedliche Anwendungskonstellation im jeweiligen Fall umgesetzt werden können.

Mit Bezug auf Überlegungen zum Geschäftsmodell für *TextTransfer* ist für den Augenblick festzuhalten, dass *TextTransfer* für die beteiligten Forschungsinstitute IDS und TIB als interner Service auf Basis einer *TextTransfer*-Pipeline vorgesehen ist. Eine Gewinnerzielungsabsicht ist hiermit nicht verbunden.

Perspektivisch verfügt die hier entwickelte Studie über vielversprechendes Potenzial hinsichtlich einer erweiterten Quellenbasis. Hierzu wäre die Methode anzupassen, um den Skopus der Impactanalyse entlang einer domänen- und quelltypspezifischen Erweiterung zu optimieren. Die im Zuge der Projektentwicklung aufgesetzte, den gesamten Workflow von Konvertierung bis maschinellem Lernen abdeckende Pipeline wäre demnach so in die Strukturen der beteiligten Forschungsinstitute zu integrieren, um dieser Zielgruppe ein Instrumentarium zur Erkennung des Impactpotenzials eigener Forschung an die Hand zu geben und so den Wissenstransfer von Forschungseinrichtungen selbst effizienter zu gestalten.

FKZ 01IO1634
Version 1.0

Zur Ausweitung des Basisprozesses hin zu einer domänen-, disziplin- und quelltypübergreifenden praktischen Anwendung ergeben sich aus den bisherigen Ergebnissen folgende Schritte der Methodenerweiterung für eine funktionale Implementierung des mit T*extTransfer* verfolgten Ansatzes in den Praxisbetrieb der Impactprognose von Forschungsinstituten:

1. Anwendungsfälle entwickeln.

Zur Schaffung einer Datengrundlage für die in *TextTransfer (Pilot)* entstandene Methode wurden ausgewählte, exemplarisch aus dem Anwendungsfeld „Mobilität“ stammende Projektberichte auf transfer- oder impactrelevante Merkmale untersucht. Im Projektverlauf zeigte sich schnell, dass für eine stabile Funktionalität der Methode neben Projektberichten auch andere Quelltypen und thematische Domänen zur Impactanalyse herangezogen werden sollten (vgl. hierzu Kap*. 3.1.1 AP 1: Bezugsrahmen* und Kap*. 3.1.2 AP2: Stichprobe*). Um die Ergiebigkeit der Fragestellung des Projekts textgestützten Transfer- oder Impactpotenzials an die auszuwählende Quellenbasis zu optimieren, sollen im Zuge der Methodenerweiterung weitere, mutmaßlich impact-affine und insbesondere gesellschaftlich relevante Themenfelder (Domänen) – zu denken wäre etwa an die im Rahmen der Hightech-Strategie der Bundesregierung identifizierten Schlüsselherausforderungen - geprüft werden. Vor dem Hintergrund der Wissenschaftsjahre 2019 mit dem Thema Künstliche Intelligenz als auch dem aktuellen, in 2020 weltweiten gesundheitlichen Notstand könnten hier auch neben dem Thema Künstliche Intelligenz Aspekte der Forschung zu gesundheitsrelevanten Themen, Demografischer Wandel, Globalisierung und Klima Gegenstand für die im Projekt anstehende Evaluation geeignete, neue und exemplarische Arbeitsfelder sein. Andererseits werden zur Optimierung der partnerspezifischen Transferfähigkeiten auch ausgewählte Felder der Linguistik als neue Domäne erschlossen werden. Darüber hinaus erscheint es sinnvoll, die bereits bearbeitete Domäne „Mobilität“ vertieft zu betrachten, um statistische Stabilität auf Basis einer erweiterten Quellenbasis zu generieren. An dieser Stelle wird die TIB ihre in *TextTransfer (Pilot)* etablierte Expertise ausbauen und die Institutionalisierung vorbereiten. Schließlich wird das IDS ebenfalls aus seinem verfügbaren Datenbestand (Berichte, Jahrbücher, Schriftenreihen, Sprachreport, wissenschaftsjournalistische Artikel in DeReKo usw.) eine Datengrundlage für die Analyse identifizieren und in Analyseprozesse überführen.

FKZ 01IO1634
Version 1.0

2. Methode erweitern.

Die in *TextTransfer (Pilot)* mittels maschinellen Lernens auf die Analyse von Projektendberichten der Domäne „Mobilität“ spezialisierte Methode soll nun auf die Anforderungen einer thematisch wie formal heterogenen Datenbasis trainiert und angepasst werden. Zu diesem Zweck werden verschieden aufbereitete und zusammengestellte Test- und Evaluierungssets aus den Stichproben dem Analyseverfahren unterzogen, um die Parameter des maschinellen Lernens weiter zu stabilisieren und zu optimieren. Die Analyseergebnisse werden kontinuierlich evaluiert und dann in das Implementierungskonzept für die technische Methode überführt.

3. Standards vorschlagen und rechtliche Rahmenbedingungen analysieren.

Zur Gewährleistung statistischer Stabilität hat die in *TextTransfer (Pilot)* entwickelte prototypische Methode vornehmlich darauf gesetzt, dass mit der dem Analyseverfahren grundlegenden Quellengattung Projektbericht Daten von hohem formalen wie inhaltlichen Standardisierungsgrad zur Verfügung standen. Schnell zeigte sich jedoch, dass zur Überführung dieser digital vorgehaltenen Texte in ein maschinenlesbares Format nichtsdestotrotz immer noch ein erheblicher technischer Aufwand zu treiben ist. Fraglos kann eine praktische Anwendung der Methode durch Veränderungen an der Form, in der Projektberichte eingereicht werden, deutlich erleichtert werden, wenn hier überarbeitete technische Standards eingehalten würden; dieser Aspekt wird zu vertiefen sein.

Die Nutzung unterschiedlicher wissenschaftlicher Quellen für Transferaktivitäten sowie die Verwendung von Text- und Datamining-Verfahren zur gezielten Auswertung solcher Quellen auf dediziert in außerakademischen Anwendungskontexten verorteten Impactpotenziale werfen überdies zahlreiche persönlichkeits-, urheber- und verwertungsrechtliche Fragen auf.

4. Implementierungskonzept ableiten.

Das in *TextTransfer (Pilot*) angestrebte Ziel war die Entwicklung einer prototypischen Methode, die es erlaubt, umfangreiche Textdaten (zunächst Projektberichte) mithilfe maschineller, linguistischer Analyseverfahren automatisiert auf Transfer- und Impactpotenzial hin zu untersuchen. An diesem Ziel orientierten sich die durchgeführten Analyse-, Recherche- und Entwicklungsarbeiten (vgl. hierzu auch Kap*. 2.1 Aufgabenstellung*). Um Anwendungsbarrieren zu überwinden, liegt es jedoch nahe, die rein technische Rohfassung für eine bedarfsgerechte Anwendung bei den Partnerinstituten

FKZ 01IO1634
Version 1.0

aufzustellen. Zentral werden hierbei die Konzeptualisierung einer technische wie organisatorische Komponenten umfassenden institutsinternen Pipeline für die Anwender im Projekt durch Absteckung geeigneter organisatorischer und anwendungsorientierter Rahmenbedingungen sein, die nach Projektabschluss auch Nutzern außerhalb des Projektkreises für ihre jeweils individuellen Anwendungsszenarien entsprechend aufbereitet zur Verfügung stehe.

5. Kommunikationskonzept entwickeln.

Neben dem gesteckten Projektrahmen ist die Sensibilisierung von und der Austausch mit Akteuren des Wissens- und Technologietransfers bezüglich der Mehrwerte der Methode *TextTransfer* von Interesse. Auch ist die Verknüpfung an ähnlich gelagerte Instrumente und Projekte zu prüfen und Erfahrungen werden auszutauschen sein. Das Projekt wird seine Ergebnisse daher so vermitteln, dass sie außerhalb des Projektrahmens bekannt gemacht werden. Es wird an dieser Stelle mit einschlägigen, vermittelnden Instanzen, etwa in der Leibniz-Gemeinschaft oder im Bedarfsfalle beim Projektträger, zusammenzuarbeiten sein. Die Ergebnisse dieser Sondierungen könnten Nachnutzungsperspektiven nach Projektabschluss eröffnen.

AP7 war zu Projektende planmäßig abgeschlossen.

3.1.8. AP8: Projektmanagement

Als Gesamtverantwortlicher hatte das IDS im Verbund sowohl die Projektleitung als auch das Gesamt-Projektmanagement inne, organisierte und koordinierte Projekttreffen als auch alle im Projekt zu erbringenden Arbeiten, sowohl der Partner als auch der Unterauftragnehmer. Das Projektmanagement innerhalb eines definierten Aufgabengebietes eines Partners lag beim jeweiligen Projektpartner.

Zu Projektbeginn trafen daher alle Projektpartner interne Maßnahmen, die einen reibungslosen Ablauf des Projektes garantieren sollten (u.a. Einrichtung einer internen Infrastruktur mit beispielsweise projektspezifischem Exchange-Server/Cloud, Zuständigkeiten etc.).

Als zentrale Kommunikationsstruktur, insbesondere zum Dokumentenaustausch der Partner, diente zu Beginn des Projektes ein virtueller Projektordner (Confluence), der vom Projektpartner TIB zur

Verfügung gestellt und administriert und vom Koordinator des IDS organisiert wurde und auf den alle Projektmitarbeiterinnen und Projektmitarbeiter Zugriff hatten.

Für den Austausch von Daten zwischen einzelnen Partnern wurde neben den Emailanwendungen, einem von der TIB eigens für *TextTransfer (Pilot)* eingerichtetem FTP-Server für den Austausch der XML-Daten zwischen der TIB und dem IDS u.a. auch für große Datenmengen Gigamove der RWTH Aachen bzw. des DFN genutzt.

Ab der zweiten Projekthälfte ersetzte zunehmend eine vom IDS zur Verfügung gestellte komfortable IDS-interne Cloud-Lösung als Austauschmedium von Dateien, insbesondere der XML-Daten, Email- als auch FTP-Anwendungen. Kollaboratives Arbeiten an Dokumenten erfolgte außerdem über Google Docs und den LaTeX Editor Overleaf.

Des Weiteren beinhaltete das Arbeitspaket die Erstellung der Dokumentationspflicht gegenüber dem Projektträger in Form von Zwischenberichten[8] als auch dem Abschlussbericht, die unter Mitarbeit aller Projektbeteiligten als auch deren Bereitstellung von Informationen und der aufgabenspezifischen Prüfung vorgenannter Berichte erstellt wurden. Die Berichte wurden dem Projektträger jeweils termingerecht vorgelegt.

AP8 war zu Projektende planmäßig abgeschlossen.

3.2. Die wichtigsten Positionen des zahlenmäßigen Nachweises

Das Projekt, dessen Laufzeit ursprünglich von 1.12.2016 bis 20.11.2018 angesetzt war, wurde - nach vorheriger Abstimmung mit dem Projektträger DLR – u.a. aufgrund den Verzögerungen bei einigen Stellenbesetzungen der Partner als auch der bei Antragstellung nicht ersichtlichen Detailkomplexität (vgl. hierzu insbesondere Kap. 3.1.2 *AP: 2 Stichprobe*) zweimal kostenneutral verlängert:

1.12.2018 bis 31.3.2019 => 1. Verlängerung (4 Monate)
1.4. bis 31.12.2019 => 2. Verlängerung (9 Monate)

[8] Zwischenberichte vom 31.12.2017 und 31.12.2018; der Zwischenbericht 2019, dessen Berichtszeitraumende mit dem Zeitpunkt des Projektendes zusammenfiel, wurde nach Absprache mit dem Projektträger in den hier vorliegenden Abschlussbericht integriert.

Alle Partner, auch die Unterauftragnehmer G&K und UIUC, sind dabei kostenneutral in die jeweiligen Verlängerungen mitgegangen (vgl. hierzu auch Kapitel 2.3.1 *Planung*).

Sämtliche Auszahlungsanordnungen für die Unterauftragnehmer G&K und UIUC des Projektpartners IDS für den ursprünglich geplanten Projektzeitraum wurden ordnungsgemäß abgerufen.

3.3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Projektträger und Projektpartner sind im Vorfeld der Anberaumung des Vorhabens *TextTransfer* zu der Erkenntnis gekommen, dass die unzähligen Ergebnisse bisherige Forschungsarbeiten aller Disziplinen in ihrer verschriftlichen Form eine wertvolle, aber bisher nicht vollumfänglich genutzte Ressource in den Archiven einschlägiger Gedächtnisorganisationen darstellen. Weiterhin stand zu erwarten, dass klassische, analoge Verfahren der Auswertung hinsichtlich verwertbarer Forschungsergebnisse nicht mehr zu ihrer Erfassung hinreichen dürften. Auf dem Wege zur Etablierung einer routinemäßigen Transferkultur in den Wissenschaften war es allen Beteiligten ein Anliegen, die Chancen der Digitalisierung auch in diesem Bereich zu nutzen. Ein automatisiertes Verfahren war zu entwickeln, das in erster Linie die Wissenschaft unterstützt, Transfer- und Impactpotenziale in wissenschaftlichen Texten besser zu identifizieren und so den Wirkungsgrad von Investitionen in die Forschung zu optimieren.

Für einen Funktionsnachweis der Methode *TextTransfer* war daher ein Zusammenspiel von einer transferrelevanten Indikatoren- bzw. Kategorienschemata-Entwicklung, der Annotation von Textquellen sowie der exemplarischen Adaption vorhandener Softwarelösungen basierend auf einen bedarfsgerecht zugeschnittenes Korpus von Forschungsberichten als Stichprobe, die im Zielformat TEI XML (i5) konvertiert sein musste, nötig.

Mit den an *TextTransfer* beteiligten Instituten und Experten hat sich ein Verbund zusammengefunden, der notwendige Kernfähigkeiten im Koprusaufbau, Text Mining,, Impact Assessment, maschinellem Lernen und Transfereigenschaften von Forschungswissen bündelt - eine Konstellation, die vor dem Hintergrund der Fragestellung im Projekt und der gewählten deutschsprachigen Datenbasis bisher nicht existierte. Eine entsprechende Förderung zur Herstellung

FKZ 01IO1634
Version 1.0

notwendiger Verknüpfungen und Kapazitäten war daher notwendig. Diese Kombination von Expertenwissen konnte durch das Projektteam IDS und TIB und den Unterauftragnehmern Görgen & Köller GmbH (G&K) und Prof. Dr. Jana Diesner von der School of Information Sciences / The *iSchool* der Universität von Illinois at Urbana-Champaign (UIUC) und ihrer Arbeitsgruppe erbracht werden. Eine solch erfolgreiche Kooperation, die aufgrund des neuartigen Ansatzes und des hohen Innovationsgrads ein hohes Forschungsrisiko birgt, wäre angesichts zu geringer personeller Ressourcen sowie der nicht hinreichend fachübergreifend breiten Kompetenzen ohne fördernde Maßnahme nicht möglich gewesen.

3.4. Voraussichtlicher Nutzen, insbesondere die Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Das Projekt *TextTransfer (Pilot)* war in seiner ersten Förderphase als Implementierungs- und Evaluierungsprojekt konzipiert, das den Funktionsnachweis für einen neuartigen Projektansatz und eine neue Methode zur Bewertung des Transfer- und Impactpotenzials von geförderten Forschungsprojekten über den akademischen Bereich hinaus erbringen sollte. Als Projektergebnis wurde der Funktionsnachweis für eine neuartige Methode maschinengestützter Analyse großer Datenmengen mit Schwerpunkt der Erkennung transfer- und impactrelevanter Eigenschaften deutschsprachiger Textdaten für den Ansatz des maschinellen Lernens mit externen Impact- und Transferkategorien (deduktiver, Top-Down-Ansatz) als auch für einen textbasierten, induktiven, Bottom-Up-Ansatz erbracht, die die bisherige Impactforschung ergänzt und neue Berechnungsmethoden zur Bewertung des Transfer- und Impactpotenzials von öffentlich finanzierten Forschungsprojekten entwickelt hat. Die Ergebnisse zeigten dabei, dass die Berichte über die finanzierten Projekte sowohl ein Potenzial für gesellschaftlichen als auch für wirtschaftlichen Impact aufweisen. Die finale Methode ist einerseits selbst inhärent als Transferinstrument entworfen und andererseits zur aktiven Nutzung bei den beteiligten Instituten angelegt.

Im Zeitalter der Herausforderung der Globalisierung durch Extremismus, Pandemien oder Krisen des Finanzsystems werden wissenschaftsgestützte Erklärungsmodelle zunehmend nachgefragt. Öffentlich gefördertes, stetig komplexer und ressourcenintensiver werdendes Wissens soll der

FKZ 01IO1634
Version 1.0

Gemeinschaft nicht durch Ablage verloren gehen. Angesichts größter Datenmengen und mit Zunahme wissenschaftlicher Arbeiten in Textform besteht ein erhöhter gesellschaftlicher Bedarf an fortschrittlichen Ansätzen zur Bewertung ihres Impacts jenseits klassisch analoger Auswertungsverfahren. Die derzeitigen Bezugsrahmen und Methoden sind arbeitsintensiv und zeitaufwendig. Nach besten Wissen des Projektteams stellt dieses Vorhaben eine der ersten Berechnungsmodelle zur Bewertung der öffentlich geförderten Transfer- und Impactforschung dar, zumal sich die Forschung zu diesem Thema, insbesondere im deutschsprachigen Raum, noch im Anfangsstadium befindet. Während kommerzielle Analyseverfahren, die zumeist auf englischsprachigen Daten basieren, an Zahl und Reichweite gewinnen, stößt *TextTransfer* explizit in die Nische bisher unerschlossener deutschsprachiger Daten. Auf diese Weise stellt sich die Methode nicht nur hinsichtlich ihres Fähigkeitsprofils, die Wahrscheinlichkeiten von Verwertungspotenzials von Forschungsergebnissen aufgrund statistischer Vergleiche zu prognostizieren, als Innovation auf. Sie leistet außerdem ihren Beitrag in der Nutzung künstlicher Intelligenz in Deutschland und öffnet nationale Kernmärkte für neue Ansätze des Impact Assessments.

Das Projekt *TextTransfer* war in seiner Pilotphase als Machbarkeitsstudie angelegt: Zum einen war die verwendete Stichprobe bezüglich des Umfangs klein. Zum anderen fand eine Konzentration auf eine Domäne (Mobilität) statt. Um dieses Manko zu beheben, müsste die Anzahl der Projekte erhöht und Berichte aus weiteren Domänen genutzt werden. Auf Grund fehlender Ressourcen, wie z.B. zusätzliche Benutzerbefragungen und Interviews, war es außerdem nicht möglich, die realen Indikatoren der im Projekt identifizierten Impactkategorien sowohl mittels des deduktiven als auch des induktiven Ansatzes abzugleichen. Erstrebenswert wäre es daher, Zugang zu mehr Informationen im Zusammenhang mit diesen Projekten zu erhalten und die Rahmenbedingungen entsprechend optimieren zu können.

In einer zweiten Förderphase für das Hauptprojekt – im Folgenden *TextTransfer* genannt – ist daher geplant, die positiven Ergebnisse von *TextTransfer (Pilot)* aufzugreifen und konsequent weiterzuführen: Kann unter Anwendung semantischer Analyseverfahren der Funktionsnachweis zur Identifizierung relevanter sprachlicher Texteigenschaften beispielhaft in deutschsprachigen

FKZ 01IO1634
Version 1.0

Forschungsprojektendberichten für die Domäne Mobilität bereits im vorliegenden Evaluierungsprojekt erbracht werden, so offenbaren sich dennoch auch Themen und Funktionalitäten, die konzeptuell erst nach erbrachtem Funktionsnachweis und daher nicht mehr im Rahmen von *TextTransfer (Pilot)* bearbeitet werden können, jedoch für eine nachhaltige Methodenanwendung unter wissenschaftlichen Realbedingungen in den beteiligten Instituten im Wissens- und Technologietransfer hohe Relevanz haben. Die Methode *TextTransfer* kann somit zu einem Instrument des Wissens- und Technologietransfers werden, das es den beteiligten Instituten auch in Zusammenarbeit mit wissenschaftlichen Dachorganisationen wie der Leibniz-Gemeinschaft erlaubt, eigene schriftliche Erzeugnisse frühzeitig und schnell nach Impactwahrscheinlichkeiten zu kategorisieren und somit die gezielte Suche nach Transfer- und Impactpotenzialen der eigenen Forschung ermöglichen. Der Schluss von Textinformationen auf Transfer- und Impactpotenzial ist mit der *TextTransfer* Methodik nicht kausal, sondern assoziativ aus beobachteten Mustern abgeleitet.

Ist die Methode bisher zum Zwecke grundsätzlicher Erprobung inhaltlich mit dem Förderbereich Mobilität als Domäne und gattungsbezogen mit Projektendberichten als Quelltyp stark fokussiert, soll aufbauend auf den ersten Vorarbeiten in der Pilotphase mit dem nächsten Schritt im Rahmen des Hauptprojektes *TextTransfer* untersucht werden, inwiefern sie auf den Einsatz mit domänenspezifisch vollkommenen unterschiedlichen Texten dieser und anderer Quelltypen übertragbar ist, um dann über eigens entwickelte prototypische Implementierungskonzepte bei den Projektpartnern zur Anwendung zu kommen. Hierfür ist geplant, dass die Projektpartner prototypische Anwendungsfälle entwickeln, um die Funktionalität der Methode unter komplexen Bedingungen der wissenschaftlichen Praxis zu demonstrieren, Wege zu ihrer Nutzung in ihrem Alltag aufzuzeigen und Anforderungen an die strukturelle Beschaffenheit wissenschaftlicher Texte mit dem Ziel einer optimierten Maschinenlesbarkeit abzuleiten. Die vorliegenden Modelle würden mit anderen populären (und qualitativen) Metriken, nämlich der Bibliometrie und der Altmetrik, kombiniert werden. Ein mögliches Hauptprojekt *TextTransfer* würde außerdem Kommunikationsmaßnahmen bereithalten, um zukünftigen, projektexternen Anwendern bereits während der laufenden Projektphase Einsicht in die Anwendungsfälle und

FKZ 01IO1634
Version 1.0

Implementierungslösungen bei den Projektpartnern zu gewähren, die eine eigene Nachnutzung anstreben.

Neben der Implementierung eines Transferinstruments bei den beteiligten Partnern, das die Identifizierungsprozesse anwendbarer Forschungsergebnisse künftig deutlich straffen und effizienter gestalten wird, vermag die Methode *TextTransfer* auch nach Projektabschluss noch die Fundamente für vielseitige wissenschaftliche wie transferbezogene Weiterentwicklungen der Partner zu legen. So ergeben sich bei der TIB etwa folgende Perspektiven einer Ausbaufähigkeit:

- Zur Verbesserung der Ausgangssituation bei künftigen Text- und Datamining-Anwendungen beruhend auf Daten aus den Deutschen Forschungsberichten wird an der TIB - unter Berücksichtigung neuester rechtlicher Grundlagen (DSGVO seit dem 25.05.2018 sowie Urheberrechts-Wissensgesellschafts-Gesetz (UrhWissG) seit dem 01.03.2018, hier insbesondere § 60d und e) - ein moderner, webbasierter Workflow zur Abgabe von Forschungsberichten entwickelt. Dieses im Rahmen von *TextTransfer* zu entwickelnde Workflow-Modell soll Berichterstatter dabei unterstützen, neben dem Bericht im PDF-Format auch strukturierte Quellformate (z.B. DOCX-Dokumente) mit einzureichen sowie die Nachnutzung unter einer geeigneten Creative-Commons-Lizenz zu erlauben. Es soll auch für andere digitale Archive und Bibliotheken einfach nachnutzbar bzw. anpassbar gestaltet sein.
- Die TIB veröffentlichte zur Vorbereitung von *TextTransfer* in Eigenleistung im 3. Quartal 2018 das prototypische Tool *Academic Document Architecture* (ADA). Dabei handelt es sich um eine auf der freien Software *Fiduswriter*[9] aufbauende Umgebung, in der browserbasiert an strukturierten Texten gearbeitet werden kann. Die Benutzerführung erinnert an das populäre „Google Texte & Tabellen“: Texte können jedoch schon bei der Eingabe semantisch angereichert werden, etwa um Informationen zur Identität der jeweiligen Urheber. Im Rahmen von *TextTransfer* entwickelt die TIB eine Dokumentenvorlage in ADA, die optional genutzt werden kann, um strukturierte, semantisch angereicherte Forschungsberichte zu erstellen und im selben Zuge einzureichen. Die TIB hat am Rande von *TextTransfer (Pilot)*

[9] vgl. u.a. https://www.fiduswriter.org/

FKZ 01IO1634
Version 1.0

eine Analyse der in den Forschungsberichten zitierten Quellen mittels automatischer Textmining-Tools vorgenommen. Basierend auf diesen Erfahrungen soll für die Bibliothek ein interaktives visuelles Dashboard auf Basis der freien Software *Kibana* entwickelt werden, indem (ohne eine initiale Suchanfrage des Benutzers vorauszusetzen) in dem Netzwerk der Urheber, Themen und Quellen von Forschungsberichten gebrowst werden kann. Dabei können auch vielfältige Ergebnisse von Analysen des IDS eingebunden werden. Ziel ist ein zeitgemäßes Interface zum intuitiven Erkennen von Zusammenhängen und Trends sowie das schrittweise Komponieren von Suchanfragen (drill down). In einem Market Place sollen Benutzer Suchanfragen anbieten und austauschen können, die in relevanten Treffermengen und Ansichten resultieren. Im Rahmen des Projektes sollen hierfür die konzeptuellen, organisatorischen und technischen Voraussetzungen ausgelegt und geschaffen werden, die künftig eine Implementierung eines solchen Tools ermöglichen.

Das IDS kann das Instrument nach Abschluss der Methodenstabilisierung in die Abläufe der Stelle für Wissenstransfer innerhalb der institutsquerschnittliche Aufgaben tragenden Abteilung *Digitale Sprachwissenschaft* und ihres Programmbereiches *Forschungskoordination und Forschungsinfrastrukturen* integrieren. Dort wird sie als Transferinstrument den Mitarbeitern des Instituts zur Verfügung stehen. Von Relevanz werden dann auch in einer möglichen Fortsetzung geleistete kommunikativen Vorarbeiten, die im Bedarfsfalle wieder aufzunehmen wären, um externe Interessenten etwa über eine bis dahin etablierte Vermittlung der Leibniz-Gemeinschaft in der Methodenanwendung zu unterstützen.

Bezüglich eigener wissenschaftlicher Anwendungsszenarien verfügt das IDS über einen großen Bestand von linguistischen Texten wie etwa Berichte und Aufsätze, für deren Analyse auf Impact und Transferpotenziale im Rahmen der *Zentralen Forschung* geeignete Strukturen für die Implementierung der entstehenden Methode vorhanden sind. So hat das IDS vor zwei Jahren dort die Position eines Transferbeauftragten dauerhaft eingerichtet. Zu dessen Aufgaben zählt auch die Identifikation von anwendbaren Forschungsergebnissen am Institut. Das Tätigkeitsprofil dieser Stelle

FKZ 01IO1634
Version 1.0

wird im Rahmen des Projekts so gestaltet werden, dass die Methode *TextTransfer*, welche das IDS als mächtiges und im Rahmen seiner Forschungskompetenzen stetig anpassungsfähiges Transferinstrument ansieht, in die Arbeitsprozesse des Institutsbereichs *Transfer* einfügen und sich individuell passende Transferketten aus den identifizierten Fällen ableiten wird. Die im Rahmen von *TextTransfer* zu entwickelnde Methode reichert den Wissenstransfer des IDS somit in drei Dimensionen an:

- Instrumentell: Neben anderen Methoden zur Identifikation des Transferpotenzials, wie z.B. Transfermatrizen, erlaubt die entstehende Methode einen automatisierten Test auf Transferierbarkeit schriftlich vorliegender Forschungsergebnisse.
- Organisatorisch: Die Aufgaben des IDS-internen Transferbeauftragten werden um den Einsatz des automatisierten Testverfahrens und die hiermit zusammenhängenden Tätigkeiten erweitert.
- Prozessual: Die bereits im IDS existierenden Transferprozesse werden, bei Vorliegen der Forschungsergebnisse und Schriftstücke in angemessener Form, um einen automatisierten Prozess ergänzt.

Sobald die Stabilität der Pipeline erwiesen ist und die Methode als Angebot der Transferstelle den Mitarbeitern des Instituts zur Verfügung steht, kann der Verarbeitungsprozess mit immer neuen Datensets – ggf. auch aus den Korpora des IDS – evaluiert werden. Auf diese Weise wird genau zu identifizieren sein, welche Komponenten in der Pipeline letztendlich generisch einsetzbar und bei welchen Komponenten potenzielle Anwender bei Anpassungen (z.B. Vorverarbeitung der Daten) zu unterstützen sein werden. Vergleichbar zum Partner IDS wird der Partner TIB den Einsatz der Methode und deren Integration in die organisatorischen Abläufe, auf die Gegebenheiten seines Hauses angepasst, vornehmen können.

Ein erfolgreicher Einsatz der Methode durch Forschungseinrichtungen setzt somit zwangsläufig die Überführung der sondierenden Pilotphase in eine praxisorientierte Hauptphase voraus.

FKZ 01IO1634
Version 1.0

3.5. Zum Zeitpunkt der Durchführung des Vorhabens dem ZE bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Es ist nicht bekannt, dass Entwicklungen anderer Einrichtungen oder Firmen die hier vorgelegte korpusgestützte Methode zur Erkennung von Verwertungsmustern in wissenschaftlichen Texten überflüssig gemacht hätten. Die im Projekt erarbeitete Methode ist derzeit einzigartig. Mit dem Projekt *TextTransfer* (Pilot) ist es nach Wissen der Beteiligten erstmals gelungen, ein Instrument zur Prognose von Verwertungspotenzialen öffentlich geförderter Forschung zu entwickeln. Auch mit Blick auf die durch die herangezogene Datengrundlage geöffnete Marktlücke deutschsprachiger wissenschaftlicher Texte stellt erschließt *TextTransfer* bisher unbearbeitetes Terrain dar.

3.6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses nach Nr. 6 (BNNest-BMBF 98)

3.6.1. Vorträge

- NORMAN FIEDLER (IDS): Vortrag „Die Maschine füttern…TextTransfer – Impact von Forschung korpuslinguistisch prognostizieren.“, AK Wissenstransfer der Leibniz-Gemeinschaft, Frankfurt a.M., 03.05.2018.
- JANA DIESNER, REZVANEH REZAPOUR (UIUC): Posterpräsentation "Assessing the Impact of Grant-Funded Research on Society by using Alternatives to Bibliometric Measures", UIUC iSchool 2018 Showcase, Urbana Champaign, 31.10.2018.
- NORMAN FIEDLER (IDS): Vortrag „Geisteswissenschaften und die neue Welt, oder: die Relevanz der Linguistik.“, Seminar: Sprachtechnologie und Sprachressourcen, Universität Mannheim, 07.11.2018.
- CHRISTOPH KÖLLER (G&K): Vortrag "Vom Ergebnis zum Produkt - Transfer sozial- und geisteswissenschaftlicher Erkenntnisse", 8. Transferwerkstatt des BMBF, Berlin, 15.11.2018.

3.6.2. Veranstaltung, Workshops, Kurse

- JANA DIESNER (UIUC), GEORG REHM (DFKI), ANDREAS WITT (IDS): Workshop " 1st Workshop on Computational Impact Detection from Text Data", LREC 2018, Miyazaki (Japan), 08.05.2018.
- Angenommen: ANDREAS WITT (IDS), JANA DIESNER (UIUC), REZVANEH REZAPOUR (UIUC): Workshop "2nd Workshop on Computational Impact Detection from Text Data", LREC 2020, Marseille (France), 11.-16.05.2020.[10]

[10] Auf Grund der weltweiten Covid 19-Pandemie fand eine analoge LREC 2020 nicht statt (vgl. https://lrec2020.lrec-conf.org/en/).

3.6.3. Publikationen/Poster

- ANDREAS WITT, JANA DIESNER, DIANA STEFFEN, REZVANEH REZAPOUR, JUTTA BOPP, NORMAN FIEDLER, CHRISTOPH KÖLLER, MANU RASTER, JENNIFER WOCKENFUSS: Paper "Impact of Scientific Research beyond Academia: An Alternative Classification Schema", LREC 2018, Miyazaki (Japan), 07.-12.05.2018.[11]
- REZVANEH REZAPOUR, JUTTA BOPP, DIANA STEFFEN, NORMAN FIEDLER, ANDREAS WITT, JANA DIESNER: Paper "Beyond Citation: Corpus-based Methods for Assessing the Impact of Research Outcomes on Society" , LREC 2020, Marseille (France), 11.-16.05.2020 [12]
- REZVANEH REZAPOUR, JUTTA BOPP, NORMAN FIEDLER, DIANA STEFFEN, ANDREAS WITT, JANA DIESNER: Poster „Beyond Citation: Corpus-based Methods for Assessing the Impact of Research Outcomes on Society." 6th Annual International Conference on Computational Social Science (IC2S2), Cambridge, MA.,17.-20.07.2020.[13]

3.6.4. Gezielte Kommunikation Projektergebnis an Fachleute

Im Rahmen der Erhebung von Daten für den deduktiven Ansatz, in dem Interviews mit Projektmitarbeitern geführt wurde (vgl. hierzu Kap. *3.1.2 AP: Stichprobe*) war es der ausdrückliche Wunsch von ca. 20% der Befragten, über die Ergebnisse in *TextTransfer (P)* nach Projektende informiert zu werden. Geplant ist daher, den Interessenten nach Freigabe des Abschlussberichtes bzw. der Verfügbarkeit in der TIB-Datenbank eine entsprechende Kommunikation durch den Projektpartner IDS zukommen zu lassen.

Mannheim, den Prof. Dr. Henning Lobin (Direktor IDS)

[11] Leibniz-Institut für Deutsche Sprache, Mannheim, Germany, University of Cologne, Germany andreas.witt@uni-koeln.de , *School of Information Sciences, University of Illinois at Urbana-Champaign, USA, *The German National Library of Science and Technology, Hannover, Germany, Görgen & Köller GmbH, Hürth, Germany (http://lrec-conf.org/workshops/lrec2018/W6/pdf/5_W6.pdf)

[12] Auf Grund der weltweiten Covid 19-Pandemie fand eine analoge LREC 2020 nicht statt (vgl. https://lrec2020.lrec-conf.org/en/). Als Konferenz-Paper angenommen, jedoch als Poster präsentiert. (http://www.lrec-conf.org/proceedings/lrec2020/index.html#6777)

[13] Auf Grund der weltweiten Covid-19-Pandemie wird die Veranstaltung virtuell stattfinden (vgl. http://2020.ic2s2.org/6th-international-conference-computational-social-science)

FKZ 01IO1634
Version 1.0

4. Anlagen

4.1. Leitfragen der telefonischen Interviews

Kurzbeschreibung *TextTransfer (Pilot)*

1. Wurde durch Ihr Projekt außerhalb der Wissenschaft ein Impact erzielt (also ein Effekt, der in anderen Bereichen der Gesellschaft und außerhalb der Wissenschaft erzielt wurde)?

2. Erfolgte dies während des Projektes oder auch noch danach?

3. In welcher Form wurde der Impact erzielt?

a) Wirtschaftlicher Impact
b) Technologischer Impact
c) Sozio-kultureller Impact
d) Politisch-rechtlicher Impact
e) Umweltbezogener und ökologischer Impact

4. Wurden durch das Projekt Erträge für Ihre Forschungseinrichtung erwirtschaftet?

5. Wünschen Sie, über die Ergebnisse von *TextTransfer (Pilot)* informiert zu werden?

FKZ 01IO1634
Version 1.0

4.2. Arbeitsanleitung für Konvertierungsarbeiten[14]

Konvertierung von PDF-Dokumenten nach TEI XML (i5) - Anleitung

Dokumente aus Sammlungen wie den „Deutschen Forschungsberichten“ an der TIB liegen oft nur im PDF-Format vor. Um sie in strukturierte Dokumente im TEI XML (i5) Format (verwendet und definiert vom IDS Mannheim) zu verwandeln, kann zunächst eine automatische Konvertierung des Dokuments nach XML 1.0 durch Adobe Acrobat Pro vorgenommen werden. Im Anschluss daran sind jedoch umfangreiche manuelle Bereinigungen erforderlich. Wie dabei vorzugehen und was im Einzelnen zu beachten ist, wird im Folgenden beschrieben. Die Fassung dieses Textes ist als Arbeitsanweisung innerhalb des BMBF-geförderten Projekts *TextTransfer (Pilot)* (12/2016-3/2019) an der TIB entstanden.

1. Übersicht: Anwendung der XML Tags

- Die Struktur der XML-Dateien basiert auf die IDS-Datei "i5.dtd".
- Meta-Informationen wie Titel, Autor, Erscheinungsdatum, etc. am Anfang des XML-Dokuments (<idsHeader>) sollten dem TIB-Katalog entnommen werden bzw. aus der PDF-Datei .

2. Grundsätzliche XML-Struktur

```
<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE idsCorpus SYSTEM "i5.dtd">

<idsCorpus>

 <idsHeader>

 <idsDoc>

   <idsHeader>

     <idsText>
```

[14] Stand März 2019 (finale Version)

FKZ 01IO1634
Version 1.0

```
<idsHeader>
<text>

<body>
  <div>
  <div>
  ...
  <div>
</body>
</text>
</idsText>
</idsDoc>
</idsCorpus>
```

3. Alle XML-Tags sollen in Kleinbuchstaben geschrieben werden, beispielsweise:

<head> …. </head>

4. Der Teil vor dem “Inhaltsverzeichnis” soll als *Titelbereich* betrachtet werden:

<head>Titelseite</head>

5. “Inhaltsverzeichnis”, “Abbildungsverzeichnis”, “Tabellenverzeichnis”, „Literaturverzeichnis“, „Berichtsblatt“ und „Document Control Sheet“ sollen jeweils als *einzelne Kapitel ( )* betrachtet werden.

6. Kapitel und Unterkapitel ergeben sich aus dem Inhalt des jeweiligen Dokuments.

FKZ 01IO1634
Version 1.0

7. Jeder -Tag enthält einen <head>-tag mit dem Name des jeweiligen (Unter)Kapitels.

Example:

<head>1. Stand der Technik </head>

2.1 Regeln für Abbildungen

8. Für *Abbildungen oder Bilder* innerhalb eines Dokumentes wird folgender XML-Tag verwendet:

<figure/>.

Der Titel einer Abbildung wird als einfacher Absatz XML-Tag dargestellt:

Abbildung 1: …….

2.2 Regeln für Absätze

9. Jede Zeile innerhalb eines -Tag wird wie ein Absatz behandelt.

…..

2.3 Regeln für Tabellen

10. In einer *Tabelle* für eine Zeilenüberschrift (fettgedruckte 1. Zeile) unterscheidet sich das Start-Tag von einem normalen Zeilen-Tag.

Beispiel:

<row role="label">

<cell> </cell>

</row>

…….

FKZ 01IO1634
Version 1.0

<row>

<cell> </cell>

</row>

Dasselbe gilt für den Tag <cell> wenn es eine Überschrift enthält (fettgedruckt).

<cell role="label"> </cell>

2.4 Regeln für Listen

Einfache Liste (mit Aufzählungszeichen:)

<list rend="bulleted">

<item></item>

……

</list>

Verschachtelte Liste (mit Aufzählungszeichen)

<list rend="bulleted">

<item>

<list rend="bulleted">

<item></item>

</list>

</item>

</list>

Nummerierte Liste

FKZ 01IO1634
Version 1.0

<list rend="numbered">

<item></item>

……

</list>

2.5 Regeln für Verweise und Fußnoten:

12. Eine *Fußnote*-Referenz wird in einem <note>-Tag eingebettet (Zahl und Text).

<note n="1">

Fußnote Text …….

</note>

2.6 Sonderseiten:

13.“*Berichtsblatt”* und “*Document Control Sheet” werden beide als einzelne Kapitel behandelt.* Ihre Strukturen unterscheiden sich ein wenig von den Standardkapiteln in einem Dokument:

1. 2.

Die Ausrichtung sollte mit dem Dokument identisch sein (zwischen den Zahlen gibt es eine Lücke).

Jede neue Absatzreihe startet mit:

3. Beschreibung der Vorgehensweise zur Bereinigung des XML-Codes

Wichtig: Es geht nicht darum, den XML-Code nur etwas anzupassen, sondern das TEI XML (i5) Format einzuhalten. Ebenfalls geht es nicht darum, das Dokument mit den verschiedenen Formatierungen wie Schriftgröße usw. anzupassen, sondern lediglich um die Rohdaten des Textes. Wir verwenden also wirklich nur die hier im Dokument genannten Tags und keine anderen! Wir erstellen ein quasi sehr vereinfacht dargestelltes Dokument in seiner Grundstruktur. Vor diesem Hintergrund sollte man

FKZ 01IO1634
Version 1.0

auf die Struktur des Original-Dokuments im PDF-Format achten und es geöffnet neben Notepad++ (oder einen anderen Texteditor, der das XML-Markup visuell hervorhebt) vor sich haben.

4. Das XML-Template

Zunächst wird vom generierten XML-Dokument aus dem PDF jeglicher XML-Code entfernt, bevor der erste Text des PDF-Dokuments zu sehen ist, der generell mit dem Titel der Arbeit beginnt. Vor diesem ersten Tag werden aus dem XML-Template des IDS die Zeilen 1-109 kopiert und in unser vorliegendes XML-Dokument eingefügt. Die Zeilen 111-115 aus dem XML-Template werden ganz am Ende des XML-Dokumentes hineinkopiert.

Um das XML-Dokument einfacher lesen zu können, fangen wir erstmal mit dem Löschen von Tags an. Dazu gehören Tags wie folgende:

<sect>; </sect>; <imagedata src=““/>, </figure>, <td/> , <figure alt…>

Umwandeln von Tags

<h1>, <h2>, <h3>, etc. in <p; dazu äquivalent </h1>, </h2>, </h3> in

<link> und </link> in und

<figure> in <figure/>

In Tabellen:

<tr> in <row>; </tr> in </row>

<th> und <td> in <cell>; </th> und </td> in </cell>

In Listen:

<lbody> in <item>; </lbody> in </item>

Sections: (Unter)Kapitel

Beim Erstellen der sections wird der öffnende Tag immer wie folgt angegeben:

FKZ 01IO1634
Version 1.0

 bzw.

Jedes (Unter)Kapitel muss einen Namen haben, z.B. "Inhaltsverzeichnis", "1. Einleitung", "2.2.3 Testreihe" etc. Der Name ist in der Rohversion meistens in einem Tag eingebettet. Dieser Tag wird in der finalen Version zu einem <head> Tag.

Beispiel:

1. Einleitung
 In diesem Projekt geht es darum .... .

wird zu

<head>1. Einleitung</head>
 In diesem Projekt geht es darum .... .

Beendet wird der Abschnitt mit

5. Tabellen

Wie oben angegeben müssen die Tabellen umgeschrieben werden. Wenn die Tabelle keine Kopfzeile hat mit darunter zugehörigen Daten, dann wird der Tag <row> verwendet und mit </row> geschlossen. Wenn eine Kopfzeile existiert z. B. etwa Kunde, Preis und darunter sind die dazugehörigen Daten eingetragen, wird der Tag <row role="label"> verwendet. Geschlossen wird die ausgezeichnete Zeile dann mit </row>.

6. Listen

Listen können nummeriert oder nicht nummeriert sein. In dem automatisch konvertierten XML-Dokument müssen praktisch immer zahlreiche Tags einfach gelöscht werden, weil diese nicht benötigt werden. Benötigt werden nur die jeweilige Listenart und der Tag <item>.

FKZ 01IO1634
Version 1.0

Im ersten Fall verwenden wir <list rend="bulleted">. Danach wird jeder einzelne Punkt mit <item> begonnen und mit </item> geschlossen. Bisweilen werden Platzhalter für Sonderzeichen (in Notepad++ oft als leeres Rechteck) an der Stelle angezeigt, wo in einer nicht-nummerierten Liste der Punkt stehen sollte. Auch diese Platzhalter sind einfach zu löschen, da durch die Auszeichnung der Liste als „bulleted" ja schon klar ist, dass es sich um eine solch punktierte Liste handelt.

Wenn die Liste nummeriert ist, wird <list rend="numbered"> verwendet und dann die jeweiligen Auflistungen auch als <item> ...</item> ausgezeichnet.

In manchen Dokumenten wird eine Liste (im Original) durch Fließtext unterbrochen und die Liste darunter fortgesetzt. Hier wird dann nicht <item> verwendet, sondern <p> und den nächsten Unterpunkt wieder als <item> gesetzt. Die Liste wird ja trotzdem weitergeführt und endet erst mit dem Tag </list>

Schema für Listen innerhalb von Listen:

<list rend="numbered">

<item>erster nummerierter Listenpunkt</item>

<item>zweiter nummerierter Listenpunkt

<list rend="bulleted">

<item>erster punktierter Listenpunkt</item>

<item>zweiter punktierter Listenpunkt</item>

</list>

</item> hier wird der zweite nummerierte Listenpunkt geschlossen, die punktierten Listenpunkte wurden dem zweiten nummerierten Listenpunkt zugeordnet

<item>dritter nummerierter Listenpunkt</item>

</list>

7. Fußnoten

Anstelle der hochgestellten Zahl für die Fußnote wird diese Zahl ersetzt mit <note n="1">. Hier wird der Text der Fußnote übernommen </note>.

8. Sonderseiten

Bei den Sonderseiten handelt es sich um die Kapitel „Berichtsblatt" und „Document Control Sheet". Sie sind normalerweise am Ende im PDF-Dokument, kommen aber auch vereinzelt am Anfang vor. In manchen Dokumenten gibt es sogar mehrere Berichtsblätter und mehrere „Document Control Sheets". Hierzu bitte den vorliegenden Beispiel-Dokumenten folgen.

Weitere Anmerkungen zur Bearbeitung

Wenn Wörter getrennt werden, dann kann es manchmal vorkommen, dass bei der Umwandlung der Getrennt-Strich beibehalten wird; dieser muss dann manuell entfernt werden.

FKZ 01IO1634
Version 1.0

4.3. Codebook für Impact Annotation – Textebene[15]

I. Impact Definition:

TextTransfer spricht dann von **Impact**, wenn durch Transfer/Verwertung angestoßene nachweisbare Effekte einer existenziellen, gesellschaftlichen, wirtschaftlichen oder ökologischen Veränderung betrachtet werden sollen. Der Impact kann dabei sowohl während der Projektlaufzeit selbst schon stattfinden (z.B. wird ein Produkt entwickelt) als auch während der Projektlaufzeit nur angestoßen werden, so dass von einem zukünftigen Impact auszugehen ist.

Anmerkung:

1. Impact kann auch die Aufrechterhaltung oder Vermeidung von Veränderungen sein, wie z.B. die Entscheidung, keine Offshore-Bohrungen in der Ostsee durchzuführen.
2. Effekte, die innerhalb des originären Forschungsbereiches und für die entsprechende fachliche Wissenschaftscommunity erzielt werden, werden hier NICHT als Impact betrachtet. (vgl. z.B. Vorträge auf wissenschaftlichen Veranstaltungen, Publikationen, studentische Bachelor- oder Masterarbeiten)

II. Impact Haupt-Kategorien mit Unterkategorien – Übersicht[16]:

- **1. Impact** auf **Domäne/Bereich/Feld**
  - Impact auf die Wirtschaft (wirt)
  - Impact auf die Umwelt (umwelt)
  - Impact auf die Gesundheit im Allgemeinen (gesund-all)
  - Impact auf Technik & Technologie (tech)
- **2. Impact** auf **Gesellschaft, öffentliche Meinung** und/oder **Wertesystem**
  - Impact auf Judikative und Legislative (legis)
  - Impact auf Gesundheitswesen als Einrichtung (gesund-sys)
  - Impact auf öffentliches Bildungswesen, (Aus)Bildung/Erziehung (bildung)
  - Impact auf Berufswelt (beruf)
  - Impact auf politische/soziale Themen (pol-sozial)
  - Impact auf Bewusstsein/Wahrnehmung (bewusst)
- **3. Impact Outcome**
  - Impact in Form von realen Produkten bzw. deren Prototypen (produkt)
  - Wissensbasierter Impact (wissen)
  - Impact in Form von Richtlinien/Guidelines (richt)
  - Sonstiger Impact (sonst)[17]

[15] Stand Dezember 2018

[16] Stand nach Adjudikation.

FKZ 01IO1634
Version 1.0

- **4. Eigenschaften/Features von Impact Outcome**
  - Neuheit (neu)
  - Sicherheit (sicher)
  - (Daten)Schutz (datenschutz)
  - Nachhaltigkeit (nachhaltig)
  - Flexibilität (flex)
  - Personalisierung (person)
  - Sonstiger Impact (sonst)[18]

III. Impact Haupt-Kategorien mit Unterkategorien – Detailbeschreibung:

	Impact Name	Impact Information	
1.	**Impact auf Domäne/Bereich/ Feld**	**Definition**	Impact auf einen genauer definierten Bereich/ein Gebiet (Anwendungsbereich, Bereich der realen Welt)
		Unterkategorien/ Erläuterung	• **Impact auf die Wirtschaft** (wirt): sowohl makroökonomische Aspekte (wirtschaftliche Entwicklungen) als auch mikroökonomische Aspekte (Vermarktung, Businessmodell, Kosten, Umsätze, Strategien) • **Impact auf die Umwelt** (umwelt): Energiewende, Umweltschutz, Nachhaltigkeit, Klimaschutz • **Impact auf die Gesundheit im Allgemeinen** (gesund-all) (*im Gegensatz zum Gesundheitssystem bei der 2. Kategorie*): mehr Verkehrssicherheit weniger Unfälle, mehr Mobilität weniger Depressionen, weniger Schadstoffe weniger Atemwegserkrankungen • **Impact auf Technik & Technologie** (tech) (z.B. Elektromobilität, automatisiertes Fahren, Hochleistungsrechner/Computertechnologie)
		Textbeispiel	• „Die Nutzung von Elektromobilität ist ein wichtiger Erfolgsfaktor, um die Unabhängigkeit von fossilen Primärenergieträgern zu erreichen und kann auch zur dynamischen Zwischenspeicherung regenerativ basierter Energiespitzen verwendet werden." (*umwelt*) • „Das Konzept des Vorhabens stützt somit das Ziel der Bundesregierung, die Nachhaltigkeit der deutschen Volkswirtschaft zu stärken und als Leitmarkt für Elektromobilität zu fungieren. " (*wirt*)

[17] Keine „finale" Unterkategorie, sie wird benutzt, wenn neue mögliche Unterkategorien von Hauptkategorie 3. entdeckt werden.
[18] Keine „finale" Unterkategorie, sie wird benutzt, wenn neue mögliche Unterkategorien von Hauptkategorie 4. entdeckt werden.

FKZ 01IO1634
Version 1.0

	Impact Name	Impact Information	
	Impact auf Gesellschaft, öffentliche Meinung und/oder Wertesystem	**Definition**	Impact auf gesellschaftliche oder öffentliche Verhältnisse, Prozesse und Institutionen; Veränderungen sozialer Normen/Regeln des gesellschaftlichen Zusammenlebens (z.B. gemeinsame Sprache, gemeinsame Normen, Regelungen für abweichendes Verhalten, gemeinsames Verständnis einer Sache)
2	**Impact auf Gesellschaft, öffentliche Meinung und/oder Wertesystem**	**Unterkategorien/ Erläuterung**	• **Impact auf Judikative und Legislative (legis): z.B. neue** Gesetze und Vorschriften, Änderungen von Gesetzen und Rechtsvorschriften, Änderungen der Auslegung etc. • **Impact auf (öffentliches/privates) Gesundheitssystem als Einrichtung (gesund-sys)** (*im Gegensatz zu Gesundheit allgemein bei der 1. Kategorie*): z.B. Impfkampagne • **Impact auf öffentliches Bildungswesen, (Aus)Bildung/Erziehung (bildung)**: z.B. "Turbo"-Abitur, neuer Master-Studiengang, Inklusion • **Impact auf Berufswelt (beruf)**: z.B. Gleichstellung der Geschlechter, Entstehen/Verschwinden von Berufen/Berufsprofilen, Arbeitslosigkeit • **Impact auf politische/soziale Themen (pol-sozial)**: z.B.: Klimawandel, Flüchtlinge/Migration, religiöse Verfolgung, Diskriminierung wegen sexueller Orientierung etc. • **Impact auf Bewusstsein/Wahrnehmung (bewusst):** z.B. Veranstaltungen (Tag der offenen Tür, Messen, Pressekonferenzen), Zeitungsartikel, Radio-/Fernsehbeiträge, Social Media, Aktionen („Hambacher Forst")
		Textbeispiel	• „Zudem profitieren Behörden und Gesetzgeber, da Empfehlungen zur sicheren Ausgestaltung von Radnabenmotoren und der Rekuperation gegeben werden." (*legis*) • „Das Ergebnis des Projektes hat gezeigt, dass ein generelles Verbot der Dieselfahrzeuge, die nicht der Euro 6-TEMP-Norm entsprechen, sich positiv auf die Erreichung der gesteckten Klimaschutzverordnung auswirkt." (*legis*) • „Die neu hinzugekommenen Hambacher Forst-Aktivisten stützen sich in ihren Protesten auf die Ergebnisse der Klimaschutz-Studie, um gegen die Abholzung durch den RWE-Konzern aufmerksam zu machen". (*bewusst*)

FKZ 01IO1634
Version 1.0

	Impact Name	**Impact Information**	
	Impact Outcome	**Definition**	Impact Outcome stellt ein finales Ergebnis bzw. Resultat eines Effektes/einer Veränderung dar.

3	**Impact Outcome**	**Unterkategorien/ Erläuterung**	• **Impact in Form von realen Produkten** bzw. **deren Prototypen** (physischer und nicht-physischer Art) **(produkt):** z.B.
			  ○ iPhone, autonomes Auto
			  ○ Apps, online Platformen, (e)Kurse, eBooks
			  ○ spezifische Daten, z.B. Listen von Emailadressen, die als Produkt von Brokern verkauft werden
			  ○ Dienstleistungen (z.B. neue Beratungs-/Serviceleistungen im Bereich der Energiewende)
			• Wissensbasierter Impact (wissen): z.B.
			  ○ Forschungsmethoden, Lern- und Lehrmethoden, Algorithmen
			  ○ Konzepte, Modelle, Daten (wenn im Bericht erwähnt, aber nicht weiterspezifiziert)
			  ○ neue/innovative effizientere technische Verfahren (die meistens zu Patenten führen)
			• Impact in Form von Richtlinien/Guidelines (richt): z.B. einheitliche Norm für Handystecker
			• **Sonstiger Impact** (**sonst**): Mögliche neue relevante Unterkategorien, die bislang keine Berücksichtigung finden

		Textbeispiel	• „Es wird eine geeignete Sicherheitsinfrastruktur konzipiert und umgesetzt, die die verschiedenen Randbedingungen der Beteiligten berücksichtigen soll.“ (*produkt*) • “Ableitung von Empfehlungen für Sicherheitsstandards für Elektrofahrzeuge basierend auf den Ergebnissen der Nutzerstudien.“ (*richt*) • „Hierzu wird ein Prognose-Modell zur Abschätzung und Bewertung der Verkehrssicherheit entwickelt.“ (*wissen*) • „Das Modell wurde im Rahmen des Vorhabens spezifiziert, entwickelt, softwaretechnisch umgesetzt sowie im Feldversuch Berlin und Dortmund validiert.“ (*produkt*)

FKZ 01IO1634
Version 1.0

	Impact Name	Impact Information	
	Eigenschaften/ Features von Impact Outcome	**Definition**	Verschiedene, besonders wichtig scheinende bzw. differenzierende Eigenschaften dienen dazu, die Kategorien 1 (Domäne), 2 (Gesellschaft) und 3 (Outcome) genauer zu spezifizieren.
4	**Eigenschaften/ Features von Impact Outcome**	**Unterkategorien/ Erläuterung**	• **Neuheit (neu)**: ein tatsächlich neuartiges Ergebnis im Sinne einer Innovation. Ergebnisse, die nur Optimierungen oder Verbesserung von bereits existierenden Ergebnissen sind, gelten nicht als neu. • **Sicherheit (sicher)**: das Ergebnis bietet/unterstützt (mehr) Sicherheit, z.B. Verkehrssicherheit, Benutzersicherheit, allgemeine Sicherheit • **(Daten)Schutz (datenschutz)**: das Ergebnis sorgt für (mehr) (Daten)Schutz, Privatsphäre • **Nachhaltigkeit (nachhaltig)**: das Ergebnis ist nachhaltig • **Flexibilität** (flex): das Ergebnis erlaubt mehr Flexibilität, z.B. von Benutzern • **Personalisierung** (person): das Ergebnis ist personalisierbar • **Sonstiger Impact** (sonst): Mögliche neue relevante Unterkategorien, die bislang keine Berücksichtigung finden
		Textbeispiel	• Der im Projekt adressierte Ansatz des Ladungsträgermanagements unter Einsatz von RFID zur Identifikation der Ladungsträger, GPS zur Ortung, ist in dieser Form noch in keinem deutschen Seehafen im Einsatz; ein hoher Innovationsgehalt ist somit gegeben. *(neu)* • "Die elektrische und funktionale Sicherheit wurde bei der Auslegung der Systemarchitektur bereits berücksichtigt." *(sicher)* • Die Daten werden bereits im Mobilfunknetz anonymisiert erfasst, so dass zu keinem Zeitpunkt ein Rückschluss auf die Rufnummer oder den Nutzer eines Mobiltelefons möglich ist. (*datenschutz*) • So ermöglichen die im FuE-Projekt realisierten Li-Ionen Energiespeicher erstmals eine praxistaugliche und wirtschaftliche Rückgewinnung und Speicherung von potentieller und kinetischer Energie (Rekuperation) z.B. für den emissionsfreien (..) Betrieb von Nutz- und Großfahrzeugen (...)Innenstädten. (*nachhaltig*) • So können Verkehrsteilnehmer bei Störfällen (...) frühzeitig auf geeignete Alternativrouten geleitet werden und somit Reisezeitverluste verringert werden. (*flex*) • Zudem berücksichtigt "immer Mobil" die individuellen (...) Einschränkungen, (...) und Präferenzen in der Verkehrsmittelnutzung der Generation 50 plus." (person)

IV. Annotationsschritte:

FKZ 01IO1634
Version 1.0

Schritt 1. Nicht alles in einem Bericht ist relevant. Bitte beachte nur die Teile, die als „relevant“ markiert sind.

Schritt 2. Bitte lies das Codebook sorgfältig durch und mach Dich mit den Impact-Kategorien, Definitionen, Erläuterungen und Beispielen (falls vorhanden) vertraut.

Schritt 3. Wir annotieren auf Satzebene

Alle Sätze, die annotiert werden sollen, sind in der Excel-Tabelle schon eingetragen (Spalte „Satz“), Wenn Kapitel und/oder Unterkapitel sich ändern, ist das auch in der Tabelle vermerkt.

Du bekommst sowohl den gesamten Bericht (pdf) als auch nur den Teil, der für die Annotation ausgewählt wurde (txt).

Schritt 4. Wähle für jeden impact-relevanten Satz die am besten geeignete Hauptkategorie aus der Dropdown-Liste in der Spalte "Kategorie". Auch wenn mehrere Kategorien in Frage kommen, sollte eine mehr Gewicht haben als andere.

4.a. Wenn ein Satz eine Impact-Kategorie enthält, diese jedoch nicht in der Tabelle aufgelistet ist, wähle bitte "Sonstiges".

4.b. Wenn zwei oder mehrere Kategorien gleichermaßen geeignet sind und Du sie nicht unterscheiden kannst, wähle "Viele" als Kategorie.

4.c Wenn ein Satz keinen Impact enthält, wähle die Kategorie „KI“ (Kein Impact).

Schritt 5. Nachdem Du die Hauptkategorie ausgewählt hast, wähle bitte die beste Unterkategorie aus und schreibe sie in die Spalte „Unterkategorie“. Einige Beispiel-Unterkategorien sind in der Tabelle aufgelistet. Du kannst sie verwenden oder „Sonstiges“ auswählen, wenn Du eine neue relevante Unterkategorie identifizierst.

5.a Wenn Du mehr als eine vorgegebene Kategorie und/oder Unterkategorie wählst oder wenn Du eine neue Unterkategorie entdeckst, markiere das mit einem „X“ in der Spalte „Check“.

FKZ 01IO1634
Version 1.0

Schritt 6. In die Spalte „Zitate“ kopierst Du die Wörter oder die Satzteile , die deine Entscheidung untermauern. Wenn die gewählte Kategorie und Unterkategorie eindeutig sind, brauchst Du in der Spalte „Notizen“ **nichts** eintragen.

6.a Wenn Du die Kategorie "Sonstiges" oder „Vieles“ gewählt hast, gib bitte in der Spalte "Notizen" weitere Informationen an.

6.b Wenn Du mehr als eine Unterkategorie gewählt hast oder wenn Du „Sonstiges“ gewählt hast, gib bitte in der Spalte "Notizen" weitere Informationen dazu an. Bei „Sonstiges“ ist es wichtig, die neue Unterkategorie zu benennen.

Allgemeine Anmerkungen:

- Jeder Satz soll isoliert betrachtet werden, d.h. in dem Satz selbst müssen konkrete Indikatoren für eine Kategorie vorhanden sein. Es reicht nicht, wenn im großen Kontext der Satz den Impact von vorherigen Sätzen „bestätigt“, ohne selbst Impact-Indikatoren zu enthalten.
- Wenn 2 oder mehrere Sätze eine „Einheit“ bilden, d.h. sie alle Indikatoren derselben Kategorie enthalten und sie sich gegenseitig verstärken, sollten sie trotzdem einzeln aufgelistet werden und nicht zusammen in derselben Zeile dargestellt werden.
- Sätze, die einen aktuellen Status beschreiben, der weder dem vorliegenden Projekt noch dessen Laufzeit in direkter Weise zugeschrieben werden kann, werden nicht annotiert.
- Inhalte von Tabellen werden nicht annotiert.
- Sätze, die Impact-Indikatoren enthalten, aber auch „Unsicherheits“-Indikatoren (z.B.„sollte“, „könnte“), sollen trotz allem annotiert werden.
- Alle Sätze, die Indikatoren auf Impact enthalten, sollen annotiert werden, auch wenn sie sich bezüglich der Wortwahl und der Satzstruktur sehr ähnlich sind.
- Auch (kleinere) Teile von möglichen (Gesamt)Ergebnissen sollen als Outcome annotiert werden, z.B. ein Motorteil einer batterie-betriebenen Maschine.

Wichtig: Besprich deine Annotation nicht mit anderen Annotatoren!

FKZ 01IO1634
Version 1.0

4.4. Textbasierte Annotation: Auswahl und Charakteristika der Berichte

Die nachfolgende Tabelle zeigt alle relevanten Informationen die textbasierte Annotation betreffend:

- Auswahl der Berichte (Berichtsname-Kürzel, Berichtsart, Autorenschaft eines Berichts/Herkunft Industriepartner und/oder Forschungspartner)
- Anzahl der zu annotierenden bzw. der zu adjudizierenden Sätze
- Angaben zu den Annotierenden (vier unterschiedliche Annotierende, sechs Paare)

#	Projekt-bericht Kürzel	Anzahl Einzelberichte pro Projekt	Berichtsart	Autorenschaft des Berichts/Herkunft Industrie (I)/ Forschung (F)	Zahl der zu annotierenden Sätze	Zahl der zu adjudizierenden Sätze	Erste(r) Annotierende(r) (*)	Zweite(r) Annotierende(r) (**)
1	Mob67a	2	Gesamtbericht	I und F	89	47	Anno1	Anno4
2	Mob70d	3	Partnerbericht	I	82	34	Anno1	Anno4
3	Mob73b	1	Partnerbericht	F	113	67	Anno2	Anno4
4	Mob121a	2	Partnerbericht	F	95	16	Anno2	Anno1
5	Mob178e	1	Gesamtbericht	I und F	62	19	Anno1	Anno2
6	Mob212d	1	Gesamtbericht	I und F	132	12	Anno1	Anno2
7	Mob219f	1	Gesamtbericht	I und F	130	6	Anno1	Anno2
8	Mob244c	2	Partnerbericht	I	95	12	Anno1	Anno2
9	Mob247c	1	Gesamtbericht	I und F	78	11	Anno1	Anno2
10	Mob270e	2	Partnerbericht	I und F	24	5	Anno1	Anno2
11	Mob302g	1	Gesamtbericht	I und F	117	6	Anno2	Anno1
12	Mob312d	1	Gesamtbericht	I und F	136	27	Anno2	Anno1
13	Mob319d	1	Gesamtbericht	I und F	36	7	Anno2	Anno1
14	Mob321b	1	Partnerbericht	F	135	11	Anno2	Anno1
15	Mob323a	1	Gesamtbericht	I und F	94	27	Anno4	Anno1
16	Mob60a	2	Partnerbericht	F	72	17	Anno4	Anno2
17	Mob66b	1	Gesamtbericht	F	91	38	Anno3	Anno1
18	Mob104c	2	Partnerbericht	F	47	16	Anno4	Anno1
19	Mob118a	1	Gesamtbericht	I und F	128	53	Anno3	Anno2
20	Mob24a	5	Partnerbericht	F	37	14	Anno1	Anno3
21	Mob50d	4	Partnerbericht	F	53	20	Anno2	Anno3
22	Mob19e	7	Partnerbericht	F	42	22	Anno1	Anno4
23	Mob36g	7	Gesamtbericht	I und F	51	33	Anno2	Anno4
24	Mob72b	4	Partnerbericht	I	53	17	Anno1	Anno2
25	Mob74a	3	Partnerbericht	I	56	33	Anno1	Anno4
26	Mob117b	2	Partnerbericht	F	51	27	Anno2	Anno4
27	Mob126a	2	Partnerbericht	I	35	16	Anno1	Anno3
28	Mob141a	3	Partnerbericht	I	51	17	Anno2	Anno3
29	Mob151c	3	Partnerbericht	I	39	17	Anno1	Anno3
30	Mob179b	3	Partnerbericht	F	39	5	Anno2	Anno3

FKZ 01IO1634
Version 1.0

#	Projekt-bericht Kürzel	Anzahl Einzelberichte pro Projekt	Berichtsart	Autorenschaft des Berichts/Herkunft Industrie (I)/ Forschung (F)	Zahl der zu annotierenden Sätze	Zahl der zu adjudizierenden Sätze	Erste(r) Annotierende(r) (*)	Zweite(r) Annotierende(r) (**)
31	Mob248b	3	Partnerbericht	I	47	29	Anno4	Anno3
32	Mob185a	3	Partnerbericht	I	53	32	Anno4	Anno1
33	Mob252f	3	Gesamtbericht	I und F	60	32	Anno4	Anno2
34	Mob278d	3	Gesamtbericht	I und F	69	14	Anno1	Anno2
35	Mob294c	4	Partnerbericht	I	48	19	Anno1	Anno4
36	Mob296a	2	Partnerbericht	I	19	14	Anno1	Anno4
37	Mob297a	2	Partnerbericht	F	80	55	Anno2	Anno4
38	Mob299a	2	Partnerbericht	F	37	11	Anno1	Anno3
39	Mob411c	3	Gesamtbericht	F	76	28	Anno3	Anno2
40	Mob201a	3	Gesamtbericht	I und F	61	21	Anno3	Anno2
41	Mob64d	4	Gesamtbericht	I und F	17	5	Anno3	Anno2
42	Mob196a	5	Partnerbericht	F	36	11	Anno3	Anno2
43	Mob145e	6	Partnerbericht	F	20	10	Anno4	Anno2
44	Mob83b	8	Partnerbericht	F	67	17	Anno1	Anno2
45	Mob103e	6	Partnerbericht	I	55	38	Anno1	Anno4
46	Mob140f	8	Partnerbericht	I	86	23	Anno2	Anno1
47	Mob147d	6	Partnerbericht	I	26	16	Anno2	Anno4
48	Mob149c	8	Partnerbericht	I	80	25	Anno2	Anno1
49	Mob156b	7	Partnerbericht	I	28	23	Anno1	Anno4
50	Mob172e	6	Gesamtbericht	I und F	67	14	Anno1	Anno2
51	Mob182g	3	Gesamtbericht	I und F	120	27	Anno2	Anno1
52	Mob197c	2	Gesamtbericht	I und F	76	9	Anno1	Anno2
53	Mob52f	9	Partnerbericht	I	81	24	Anno3	Anno2
54	Mob56i	10	Partnerbericht	I	72	42	Anno3	Anno4
55	Mob57c	7	Partnerbericht	F	70	49	Anno3	Anno4
56	Mob59f	7	Partnerbericht	I	42	27	Anno3	Anno4
57	Mob68a	7	Gesamtbericht	I und F	89	22	Anno3	Anno1
58	Mob164d	5	Partnerbericht	F	66	43	Anno1	Anno4
59	Mob173f	5	Gesamtbericht	I und F	56	17	Anno1	Anno4
60	Mob193e	5	Gesamtbericht	I und F	29	20	Anno2	Anno4
61	Mob207e	5	Gesamtbericht	I und F	104	16	Anno2	Anno3
62	Mob100i	7	Gesamtbericht	I und F	66	34	Anno2	Anno4
63	Mob101e	5	Partnerbericht	F	126	14	Anno2	Anno3
64	Mob106b	5	Partnerbericht	F	95	21	Anno2	Anno1
65	Mob208d	3	Gesamtbericht	I und F	88	36	Anno3	Anno1
66	Mob211c	3	Gesamtbericht	I und F	30	17	Anno3	Anno1
67	Mob220a	4	Partnerbericht	I	92	48	Anno4	Anno1
68	Mob231f	3	Gesamtbericht	I und F	57	37	Anno2	Anno4
69	Mob254c	5	Partnerbericht	I	60	40	Anno4	Anno2
70	Mob264g	5	Partnerbericht	F	42	7	Anno2	Anno1

#	Projekt-bericht Kürzel	Anzahl Einzelberichte pro Projekt	Berichtsart	Autorenschaft des Berichts/Herkunft Industrie (I)/ Forschung (F)	Zahl der zu annotierenden Sätze	Zahl der zu adjudizierenden Sätze	Erste(r) Annotierende(r) (*)	Zweite(r) Annotierende(r) (**)
71	Mob272d	4	Gesamtbericht	I	33	17	Anno4	Anno1
72	Mob301d	5	Gesamtbericht	I und F	93	49	Anno1	Anno4
73	Mob313c	5	Partnerbericht	I	131	45	Anno2	Anno3
74	Mob314d	5	Partnerbericht	I	45	8	Anno2	Anno3
75	Mob320b	5	Partnerbericht	I und F	154	30	Anno1	Anno3
76	Mob329e	5	Partnerbericht	F	41	8	Anno2	Anno3
77	Mob333d	4	Partnerbericht	I	55	17	Anno2	Anno3
78	Mob335d	4	Partnerbericht	F	120	74	Anno1	Anno4
79	Mob227f	6	Gesamtbericht	I und F	96	63	Anno4	Anno2
80	Mob232h	8	Partnerbericht	I	59	46	Anno4	Anno2
81	Mob236a	5	Gesamtbericht	I und F	94	69	Anno4	Anno1
82	Mob238e	5	Gesamtbericht	I und F	51	49	Anno4	Anno2
83	Mob77b	7	Partnerbericht	I	63	43	Anno2	Anno4
84	Mob163g	6	Gesamtbericht	I und F	51	38	Anno1	Anno4
85	Mob246h	8	Gesamtbericht	I und F	57	42	Anno2	Anno4
86	Mob267h	7	Gesamtbericht	I und F	108	79	Anno1	Anno4
87	Mob215f	6	Gesamtbericht	I und F	106	77	Anno2	Anno4
88	Mob225g	6	Partnerbericht	F	97	63	Anno1	Anno4
89	Mob84e	8	Partnerbericht	F	60	44	Anno2	Anno4
90	Mob150d	8	Partnerbericht	I	45	36	Anno2	Anno4
91	Mob206g	7	Gesamtbericht	I und F	69	40	Anno2	Anno4
					6.384	**2.576**		
					100%	40%		

*Annotierende		**Annotierende-Paare	
Anno1	JA	Anno1	Anno4
Anno2	MG	Anno2	Anno4
Anno3	SVS	Anno1	Anno3
Anno4	SF	Anno3	Anno2
		Anno2	Anno1
		Anno3	Anno4

FKZ 01IO1634
Version 1.0

4.5. *TextTransfer* Project-Pipeline

developed by: Rezvaneh (Shadi) Rezapour (rezapou2@illinois.edu)

Co-PI: Jana Diesner (jdiesner@illinois.edu)

University of Illinois at Urbana-Champaign

Date: 2019/09/10

==========

Phase 1:

Codes:

1. Concat texts.py -> Concatenate the documents in each project folder to create one text file for each project.

2. NLP TextAnalysis.ipynb -> Preprocess and analyze the texts

3. ML Phase1-Unigram 0801.ipynb -> the Machine learning model for the Unigram Bag of word model

4. ML Phase1-Ngram 0801.ipynb -> the Machine learning model for the Ngram (n=3) Bag of word model

5. ML Phase1-Ngram+POS 0801.ipynb -> the Machine learning model for the Ngram Bag of word model in addition to the Part of Speech features

6. ML Phase1-Ngram+POS+Sub-Category 0801.ipynb -> the Machine learning model for the Ngram Bag of word model in addition to the Part of Speech features and the sub-categories

How to run the model:

Requirements:

Python 3,

FKZ 01IO1634
Version 1.0

Jupyter Notebook

Libraries:

Sklearn, TextBlob (the German package: textblob de), NLTK (German)

Steps:

1- Run the "Concat texts.py" code to concatenate the files in each project. Don't forget to modify the paths as highlighted in the code.

2- Run the "NLP TextAnalysis.ipynb" to preprocess the texts, and get the pos tags of each file. Modify the input and output paths and save the pickle file (which contains the preprocessed texts) as well as the csv file for the pos features.

3- Run the machine learning models, to get the prediction accuracy of each model. Don't forget to modify the paths as highlighted in the texts.

==========

Phase 2:

Codes:

1. ML Phase2 Unigram 08012019.ipynb-> the Machine learning model for the Unigram Bag of word model

2. ML Phase2 Ngram 08012019.ipynb -> the Machine learning model for the Ngram (n=3) Bag of word model

3. ML Phase2 Ngram+POS 08012019.ipynb -> the Machine learning model for the Ngram Bag of word model in addition to the Part of Speech features

4. 4)ML Phase2 Ngram+POS+Sub-Category 08012019.ipynb -> the Machine learning model for the Ngram Bag of word model in addition to the Part of Speech features and the sub-categories

FKZ 01IO1634
Version 1.0

How to run the model:

Requirements:

Python 3,

Jupyter Notebook

Libraries:

Sklearn, TextBlob (the German package: textblob de), NLTK (German)

Steps:

1- The text analysis is performed in each of the machine learning models.

2- Run the machine learning models, to get the prediction accuracy of each model. Don't forget to modify the paths as highlighted in the texts.

FKZ 01IO1634
Version 1.0

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht
3. Titel TextTransfer (Pilot): Korpusgestützte Erkennung von Verwertungsmustern in wissenschaftlichen Texten Teilprojekt IDS: Analysemethoden und Anwendungsfälle	
4. Autor(en) [Name(n), Vorname(n)] Prof. Dr. Andreas Witt Abteilungsleiter Digitale Sprachwissenschaft, IDS	5. Abschlussdatum des Vorhabens 31.12.2019 6. Veröffentlichungsdatum: 30.06.2020 7. Form der Publikation: Abschlussbericht
8. Durchführende Institution(en) (Name, Adresse) Leibniz-Institut für Deutsche Sprache (IDS) Postfach 101621, 68016 Mannheim https://www1.ids-mannheim.de	9. Ber. Nr. Durchführende Institution 10. Förderkennzeichen: 01IO1634 11. Seitenzahl: 86
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF), 53170 Bonn	13. Literaturangaben: 48 14. Tabellen: - 15. Abbildungen: 14
16. Zusätzliche Angaben	

17. Vorgelegt bei (Titel, Ort, Datum): Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) I Projektträger im DLR I Bereich Gesellschaft, Innovation, Technologie I Heinrich-Konen-Straße 1 I 53227 Bonn (Juni 2020)
18. Kurzfassung Zielsetzung: Die zentrale Aufgabenstellung des Verbundprojektes *TextTransfer (Pilot)* der Projektpartner IDS und TIB war eine Machbarkeitsprüfung für die Entwicklung eines Text-Mining-Verfahrens, mit dem Forschungsergebnisse automatisiert auf Hinweise zu Transfer- und Impactpotenzialen untersucht werden können. Methode: Nach Festlegung des Bezugsrahmens – Dokumententyp Projektendberichte aus der Domäne Mobilität – wurde eine Grundgesamtheit generiert, aus der eine geeignete, maschinenlesbare Datenbasis einer in das maschinenlesbare Format TEI XML (i5) konvertierten Stichprobe zu ziehen war. Der Projektansatz bestand darin, für das überwachte Lernverfahren (supervised machine learning) zunächst ein Trainingsdatenset aus der Stichprobe auszuwählen, das mit bestimmten impactrelevante Texteigenschaften repräsentierenden Informationen angereichert wurde, so dass die Maschine auf vorgegebene Zusammenhänge von Texteigenschaften und Impact-Klassifikationen trainiert werden konnte. In einem zweiten Schritt wurden dann Indizien für ähnliche Zusammenhänge in der Maschine unbekannten, nicht vorab klassifizierten Textmengen (Evaluationsset) gesucht (distant reading). Ergebnis: Eine auf maschinellem Lernen basierende Methode wurde entwickelt, die mittels statistischem, textstrukturellem Vergleich sprachlicher Ähnlichkeiten in Projektendberichten automatisiert Transfer- und Impact-Wahrscheinlichkeiten nachzuweisen in der Lage ist. Sowohl mit Blick auf diese Fertigkeit als auch auf die eigens erstellte, komplexe Datengrundlage stellt das Vorhaben ein Unikat sowie ein Novum dar, insbesondere mit Bezug auf den deutschen Sprachraum.
19. Schlagwörter Text-Mining, Maschinelles Lernen, Korpusanalyse, Computerlinguistik, Korpuslinguistik, Impact, Impact-Assessment, Wissenstransfer, Forschungsimpact, Impact-Indikatoren, Transfer-Potenzial, IDS, TIB

20. Verlag	21. Preis

Document Control Sheet

1. ISBN or ISSN	2. type of document (e.g. report, publication) Final joint project report
3. title TextTransfer – Corpus based detection of secondary practical usage of scientific publications Subproject IDS: Analysis methods and use cases	
4. author(s) (family name, first name(s)) Prof. Dr. Andreas Witt Department Head Digital Linguistics, IDS	5. end of project: 31.12.2029 6. publication date: 30.06.2020 7. form of publication: Final report
8. performing organization(s) (name, address) Leibniz-Institut für Deutsche Sprache (IDS) Postfach 101621, 68016 Mannheim https://www1.ids-mannheim.de	9. originator's report no.: 01IO1634 10. reference no. 11. no. of pages: 86
12. sponsoring agency (name, address) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. no. of references: : 48 14. no. of tables: : - 15. no. of figures: 14
16. supplementary notes	
17. presented at (title, place, date) Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) I Projektträger im DLR I Bereich Gesellschaft, Innovation, Technologie I Heinrich-Konen-Straße 1 I 53227 Bonn (Juni 2020)	

18. abstract

Objective:

The central task of the joint project *TextTransfer (Pilot)* of the project partners IDS and TIB was a feasibility study for the development of a text-mining procedure, with which research results can be automatically examined for indications of transfer and impact potentials.

Method:

After defining the reference framework - document type final project reports from the mobility domain - a basic sample was generated from which a suitable, machine-readable database of a final sample converted into the machine-readable format TEI XML (i5) was to be drawn. The project approach consisted in selecting a training data set from the sample for the supervised machine learning process. The training data set was enriched with information representing certain impact-relevant text properties, so that the machine could be trained for predefined correlations of text properties and impact classifications. In a second step, evidence for similar correlations was then searched for in the machine unknown, not previously classified text sets (evaluation set) (distant reading).

Result:

A method based on machine learning was developed, which is able to automatically shows transfer and impact potentials by means of statistical, text-structural comparison of linguistic similarities in final project reports. With regard to this skill as well as the complex data basis created in-house, the project is both unique and a novelty, especially with regard to the German language area.

19. keywords

text-mining, machine learning, corpus analysis, computer/corpus linguistics, impact, impact assessment, knowledge transfer, research impact, impact indicators, transfer potential, IDS, TIB

20. publisher	21. price