IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

Verbundprojekt

TextTransfer II (Hauptprojekt)

Methode zur korpusgestützten Prognose von Impactmustern in wissenschaftlichen Texten

Abschlussbericht Gesamtprojekt

nach Nr. 3.2. BNBest-BMBF 98

eingereicht durch Leibniz-Institut für Deutsche Sprache (IDS), Mannheim

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

TEXTTRANSFER

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

1. Eckdaten des Forschungsprojektes

Vorhabenbezeichnung	*TextTransferII (Hauptprojekt)* – Methode zur korpusgestützten Prognose von Impactmustern in wissenschaftlichen Texten
Projektart	Verbundprojekt
Verbundpartner	▪ Leibniz-Institut für Deutsche Sprache (IDS), Mannheim *Teilprojekt IDS: Implementierung und Praxis im Anwendungsfall Linguistik* ▪ Technische Informationsbibliothek (TIB) - Leibniz-Informationszentrum Technik und Naturwissenschaft, Universitätsbibliothek, Hannover *Teilprojekt TIB: Implementierung und Praxis im Anwendungsfall Bibliothekswesen*
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	▪ 01IO2002A (IDS) ▪ 01IO2002B (TIB)
Förderer	Bundesministerium für Bildung und Forschung (BMBF)
Projektträger	01.06.2020-30.11.2023: Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR); Bereich Gesellschaft, Innovation, Technologie Heinrich-Konen-Straße 1, 53227 Bonn 01.12.2023-31.05.2024: Projektträger Jülich; Nachhaltige Entwicklung und Innovation; Hochschulen, Innovationsstrukturen, Gesundheit I Hochschulen (HIG 5); Forschungszentrum Jülich GmbH, 52425 Jülich
Laufzeit des Vorhabens	01.06.2020 – 31.05.2023 (TIB) 01.06.2020 – 31.05.2024 (IDS, inkl. Kostenneutraler Laufzeitverlängerung)
Berichtszeitraum	01.06.2020 – 31.05.2024
Berichtstyp	Gesamtabschlussbericht

Stichwörter	Text Mining, Maschinelles Lernen, Korpusanalyse, Computerlinguistik, Große Sprachmodelle, Large Language Models, LLM, Korpuslinguistik, Impact, Impact Assessment, Wissenstransfer, Forschungsimpact, Impact-Indikatoren, Transfer-Potenzial, IDS, TIB
Version	1.1
Verfasser	Prof. Dr. Andreas Witt (IDS)
Datum	30.11.2024

Inhalt:

Abbildungsverzeichnis:

FKZ 01IO2002A
Version 1.1

2. Kurze Darstellung

2.1. Aufgabenstellung

Forschungsergebnisse liegen zumeist als wissenschaftliche Texterzeugnisse vor, deren unüberschaubare Vielzahl in ihrem Gehalt und ihren Anwendungsmöglichkeiten oft nur zeitverzögert oder gar nicht wahrgenommen wird. Gleichzeitig wird – trotz der Investitionen öffentlicher Mittel in die Langzeitarchivierung – das in diesen Quellen vermutete Potenzial aus der Anwendung von Forschungsergebnissen bislang nicht hinreichend genutzt. Die Ausschöpfung des gesellschaftlichen oder wirtschaftlichen Potenzials aus der öffentlichen Rezeption und Anwendung von Forschungsergebnissen kann jedoch von wissenschaftlichen Publikationen mangels gemeinsamer Sprache und fehlender Kompatibilität in den zugrundeliegenden Formaten kaum geleistet werden. Der Forschung selbst fehlen trotz vorhandener Strukturen der Portfolio- und Marktanalyse bisweilen eigene Fähigkeiten und Kapazitäten, ihre meist schriftlich fixierten Forschungsergebnisse zeitökonomisch und zuverlässig auf erweiterte Anwendungspotenziale außerhalb der Wissenschaft zu untersuchen.

Entsprechend komplex gestaltet es sich, mit vertretbarem Aufwand und ohne allzu umfassende Fachkenntnis einzuschätzen, inwieweit unübersehbare Datenmengen wissenschaftlicher Texte hinsichtlich sprachlicher Hinweise auf durch bestimmte Formate oder Prozesse der Forschung (Wissens- und Technologietransfer bis hin zu wirtschaftlicher Verwertung) angestoßene Effekte (Impact) einer existenziellen, gesellschaftlichen, wirtschaftlichen oder ökologischen Veränderung untersucht werden können, um das Potenzial gesellschaftlicher Investitionen in problemorientierte Forschung noch besser ausschöpfen zu können als bisher.

Klassische Auswertungsverfahren sind hierfür überfordert. Einfache Wortsuchen in mutmaßlich relevanten Texten führen zumeist durch einen zu heterogenen Sprachduktus schon innerhalb derselben Disziplin kaum zu brauchbaren Ergebnissen. Darüber hinaus eröffnen sie nur selten bzw. nur mit erheblichem Aufwand einen Sinnzusammenhang zwischen Projektgegenstand, angewandter Forschungsmethode und ggf. transferrelevantem Ergebnis. Wesentliche Messgröße für solche Verfahren wäre hiernach der Impact von Forschung, der für die folgenden Ausführungen als empirisch messbare

FKZ 01IO2002A
Version 1.1

Auswirkung oder Veränderung auf bzw. Nutzen für Wirtschaft, Gesellschaft, Politik/Recht, Technologie oder Umwelt verstanden wird. Wesentliche Ursache solcher Effekte wäre die Anwendung wissenschaftlicher Forschungsergebnisse insbesondere über den akademischen Bereich hinaus. In diesem Spannungsfeld hat das Projekt in einer ersten Förderphase ((FKZ: 01IO1634 (IDS) / FKZ: 01IO1635 (TIB), im Folgenden *TextTransfer I (Pilot)*)[1] genannt, erfolgreich den Beweis erbracht, mittels eines technischen Analyseverfahrens automatisiert in großen Berichtsdatenmengen regelhafte sprachliche Merkmale zu identifizieren, die Projekten mit hohem Transfer- und Impactpotenzial (Wahrscheinlichkeiten aufgrund textstruktureller oder transferbezogener indikatorgestützter Elemente) eigen sind. Das hier vorliegende Hauptprojekt *TextTransfer II* griff die Ergebnisse von *TextTransfer I* auf und führte sie konsequent weiter.

Die zentrale Aufgabenstellung des Verbundprojektes *TextTransfer II* sah dabei in seiner Zielplanung die Stabilisierung der in der Pilotphase entwickelten Methode mittels Erweiterung der Datenbasis um neue Quelltypen und Domänen vor, um die Reichweite des entwickelten Transferinstruments je nach Nutzungsart skalierbar zu gestalten. Darüber hinaus wurden weiterführende Referenzindikatoren und Methoden zur automatischen Rekonstruktion von Wirkweisen rezipierten Wissens geprüft.

2.2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Die vom Projektgegenstand vorgegebenen einzelnen Arbeitsschritte und Aufgaben haben sich aus den spezifischen Kompetenzen der Projektbeteiligten abgeleitet. Für die Bearbeitung deutschsprachiger Textdaten hinsichtlich der im Projekt gewählten Fragestellung wurde erneut auf die im Pilotprojekt von *TextTransfer* etablierten Strukturen zurückgegriffen – mit den vorhandenen Kompetenzen im Bereich Korpusaufbau, Text Mining und maschinellem Impact Assessment.

Die Aufgaben der Verbundpartner stellten sich dabei wie folgt dar:

[1] Der Abschlussbericht von *TextTransfer I* findet sich in der TIB Bibliothek öffentlich zugänglich unter https://www.tib.eu/de/suchen/id/TIBKAT:1747197327/TextTransfer-Pilot-korpusgestützte-Erkennung-von?cHash=f56c7df1117392f268358ce611858ca4

Der Projektkoordinator **IDS**[2] konzentrierte sich im Rahmen von *TextTransfer II* auf die Auswahl der Stichprobe aus verschiedenen Domänen und Quelltypen, die Vorbereitung der Daten für das maschinelle Lernverfahren (die Annotation von Textquellen) als auch an Empfehlungen zur Methodenimplementierung hinsichtlich der technischen und rechtlichen Rahmenbedingungen gemeinsam mit dem Projektpartner TIB. Dem IDS oblagen überdies sämtliche projektadministrative Aufgaben.

Die Aufgabe des Projektpartners **TIB**[3] war es, ein bedarfsgerecht zugeschnittenes Korpus von Forschungsberichten als Stichprobe im PDF-Format zur Verfügung zu stellen. Die TIB gewährleistete außerdem die automatische Konvertierung der Stichprobe in das Zielformat txt, das zur Vorbereitung der Daten für das Maschinelle Lernen notwendig war. Die TIB arbeitete außerdem gemeinsam mit dem Projektpartner IDS an Empfehlungen zur Methodenimplementierung hinsichtlich der technischen und rechtlichen Rahmenbedingungen.

Unterstützt wurden die Verbundpartner von den Unterauftragsnehmern **Görgen & Köller GmbH (G&K)**[4] und **Prof. Dr. Jana Diesner** von der **School of Information Sciences / The *iSchool* der Universität von Illinois at Urbana-Champaign** (**UIUC**)[5] und ihrer Arbeitsgruppe. G&K unterstützte das IDS bei der Erbringung von Transfernachweisen für die die Stichprobe bildenden Projekte bzw. sowohl beim Implementierungs- als auch Kommunikationskonzept. Die UIUC passte die in *TextTransfer I (Pilot)* zur Textanalyse notwendige Software entsprechend der neu hinzugekommenen Domänen und Quelltypen an.

[2] http://www1.ids-mannheim.de
[3] https://www.tib.eu/de/
[4] https://gk-mb.com
[5] https://ischool.illinois.edu/people/jana-diesner

2.3. Planung und Ablauf des Vorhabens

2.3.1. Planung

TextTransfer hatte bereits im Rahmen einer ersten Pilotphase[6] erfolgreich gezeigt, wie unter Verwendung linguistischer Methoden der Korpusanalyse und des maschinellen Lernens eine automatisierte Auswertung exemplarisch anhand von Projektberichten auf der Suche nach bestimmtes Impactpotenzial indizierenden sprachlichen Mustern und Zusammenhängen prinzipiell zu bewerkstelligen ist.

Mit seinen auf Teilprojektebene angesiedelten Einzelkomponenten konnte die Vorstudie *TextTransfer I (Pilot)* damit 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 war nicht zuletzt auf personelle Engpässe qualifizierter Computerlinguisten und Informatiker im wissenschaftlichen Umfeld zurückzuführen. Die Knappheiten auf dem Arbeitsmarkt für qualifizierte Informatiker und der durch die Corona-Pandemie erzeugte zusätzlich sich auswirkende Rückstand in der Ausbildung des wissenschaftlichen Nachwuchses trugen dazu bei. Mit *TextTransfer I* lag eine prototypische Methode vor, die aus Textquellen im statistischen Vergleich zu einer breiten Basis an Präzedenzen Impactwahrscheinlichkeiten von Forschungsendberichten zum Thema Mobilität stabil erschließen kann. Auf dieser Grundlage ergaben sich zahlreiche Anknüpfungspunkte der Methodenstabilisierung hinsichtlich Prognosepräzision und Skalierbarkeit im Hauptprojekt. Das Hauptprojekt sah dabei einen Ausbau der bestehenden Datenbasis und die Hinzuziehung neuer Quelltypen vor.

Pandemiebedingt unterlagen beide Projektpartner von Frühjahr 2020 bis Sommer 2022 teilweise starken Betriebseinschränkungen bis hin zur Schließung. Die Rahmenbedingungen am Arbeitsmarkt sowie die starken Mobilitätseinschränkungen für wissenschaftliche Mitarbeiter in Zeiten der Corona-Pandemie haben überdies dafür gesorgt, dass der wissenschaftliche Personalbestand des Vorhabens unvorhersehbar eingebrochen war.

[6] Vgl. hierzu das Pilotprojekt *TextTransfer* 2016-2019 (FKZ 01IO1634 (IDS); 01IO1635 (TIB))

Auf Grund des unvorhergesehenen Projektverlaufs kamen der Projektträger als auch die Projektleitung des IDS bereits im Jahr 2021 überein, im Anschluss an die offizielle Projektlaufzeit in eine kostenneutrale Verlängerung zu gehen, um den Erfolg des Projektes nicht zu gefährden. Das Projekt wurde damit offiziell ein Jahr nach dem ursprünglich geplanten Termin am 31.05.2024 erfolgreich beendet. Sowohl der Projektpartner TIB, der seine Arbeiten im Projekt zum ursprünglich vorgesehenen Termin 31.05.2023 beendet hatte, als auch die beiden Unterauftragnehmer waren dennoch bis zum Ende des Gesamtprojektes unterstützend am Verbundprojekt beteiligt.

2.3.2. Abstimmung innerhalb von *TextTransfer II*

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.

Statustreffen (virtuell oder in Präsens) des Projektteams fanden zu folgenden Terminen statt:

- #1 21.04.2021 / virtuell,
- #2 16.06.2021 / virtuell,
- #3 19.08.2021 / virtuelles Meeting geplant, musste kurzfristig abgesagt werden,
- #4 09.11.2021 / Präsens-Treffen in den Räumlichkeiten des Projektpartners TIB / Hannover,
- #5 06.12.2021 / virtuell,
- #6 22.02.2022 / virtuell,
- #7 07.04.2022 / virtuell,
- #8 30.06.2022 / virtuell,
- #9 03.11.2022 / virtuell,
- #11 18.07.2023 / virtuell,
- #12 09.10.2023 / virtuell,
- #13 14.12.2023 / virtuell.

Präsenztreffen des Projektteams unter Teilnahme u.a. des Beirats fanden zu folgenden Terminen statt:

- #10 22.03.2023 / Präsenstreffen in den Räumlichkeiten der TIB, Hannover,
- #14 28.05.2024 / Projektabschlusstreffen in Präsens in den Räumlichkeiten des IDS, Mannheim.

Zwecks Absprachen und Zusammenarbeit fanden bei Bedarf außerdem Treffen – virtuell oder in Präsenz – in kleineren Gruppen statt.

Darüber hinaus wurde der Projektstatus Quo bei für das Gesamtprojekt relevanten Entwicklungen zeitnah per Cloudlösung, im Rahmen eines gemeinsamen Rolling Documents 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 Projekt *"TextTransfer II (Hauptprojekt)* – Methode zur korpusgestützten Prognose von Impactmustern in wissenschaftlichen Texten" wurde vom Bundesministerium für Bildung und Forschung (BMBF) unter den Förderkennzahlen 01IO2002A (IDS) und 01IO2002B (TIB) gefördert und vom Deutschen Zentrum für Luft- und Raumfahrt e.V. (DLR) vom 01.06.2020 bis zum 30.11.2023 und vom Projektträger Jülich – Nachhaltige Entwicklung und Innovation; Hochschulen, Innovationsstrukturen, Gesundheit vom 01.12.2023 bis 31.05.2024 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 Verlängerung 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., Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. SAGE Open, 9(1):2158244019829575.

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

Becker, D. R., Harris, C. C., McLaughlin, W. J., 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., Shin, H. (2016). Changing appetites & changing minds: Measuring the impact of" food, inc.". The USC Annenberg Norman Lear Center.

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

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. 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., Kegelmeyer, W. P. (2002). Smote: synthetic minority oversampling technique. Journal of artificial intelligence research, 16:321–357.

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

Coates Ulrichsen, T., Athanassopoulou, N. (2024). Commercialising Social Science Research: Insights from the University of Cambridge on key barriers, enablers and pathways to success. IfM Engage and the Policy Evidence Unit for University Commercialisation and Innovation, University of Cambridge. https://doi.org/10.17863/CAM.107657

Devlin, J., Chang, M.-W., Lee K., Toutanova, K.. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

FKZ 01IO2002A
Version 1.1

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

Gabrys-Deutscher, E., Lütjen, A. (2022). Deutsche Forschungsberichte – eine Sonderform Grauer Literatur in der Technischen Informationsbibliothek (TIB): Rückblick auf über ein halbes Jahrhundert Bestandsaufbau und Ausblick in die Zukunft. O-Bib. Das Offene Bibliotheksjournal Herausgeber VDB, 9(1), 1-13. https://doi.org/10.5282/o-bib/5768

Gomes, D. 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.

Gori, M., 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., Glover, M. (2016). Research impact: a narrative review. BMC medicine, 14(1):78.

Grimmer, J., 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.

Han, K., Rezapour, R., Nakamura, K., Devkota, D., Miller, D.C., Jana Diesner. 2023. An expert-in-the-loop method for domain-specific document categorization based on small training data. Journal of the Association for Information Science and Technology, 74(6):669–684.

Hennon, D., Dewaele, A., De Smet, E., Buysse, A. (2019): Guide to Impact Planning, Ghent 2019.

Heyeres, M., Tsey, K., Yang, Y., Yan, L., 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., 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.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

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

Ma, Y., 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., Uzzi, B. (2019). Women who win prizes get less money and prestige.

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

Otto, M., Scherer, A. (2015): Technologietransfer in eigenständigen Organisationsformen: ein Leitfaden für die außeruniversitäre Forschung, Potsdam: Deutsches GeoForschungsZentrum GFZ, 57 S. (DOI: http://doi.org/10.2312/GFZ.WTT.2015.001)

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

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., 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., 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.

Schomaker, Rahel M.; Sitter, A. (2020): Die PESTEL-Analyse – Status quo und innovative Anpassungen. Der Betriebswirt 61(1): 9-27. Weblink:https://www.researchgate.net/profile/Rahel-Schomaker/publication/341876583_Die_PESTEL-Analyse_-_Status_quo_und_innovative_Anpassungen/links/5fd6285192851c13fe816690/Die-PESTEL-Analyse-Status-quo-und-innovative-Anpassungen.pdf

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

Smalheiser, N. R., 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., 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., 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.

Vinkler, P. 2010. The evaluation of research by scientometric indicators. Elsevier.

Williams, K., Jonathan Grant. 2018. A comparative review of how the policy and procedures to assess research impact evolved in Australia and the UK. Research Evaluation, 27(2):93–105.

Witt, A., Diesner, J., Steffen, D., Rezapour, R., Bopp, J., Fiedler, N., Köller, C., Raster, M., 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., 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., 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/

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.

FKZ 01IO2002A
Version 1.1

3. Eingehende Darstellung

3.1. Verwendung der Zuwendung und der erzielten Ergebnisse im Einzelnen

Bei Beendigung des Projektes waren die wissenschaftlich-technischen Ergebnisse erreicht. Nachfolgend finden sich die Ergebnisse aus den Arbeitspaketen ausführlich dargestellt. Die im Folgenden dargestellten Arbeitspakete (AP) bilden die einzelnen Arbeitsschritte des Vorhabens *TextTransfer II* ab. In seinem grundsätzlichen Vorgehen baut es auf den Erkenntnissen des Pilotprojektes auf, das den prinzipiellen Funktionsnachweis der Methode erbracht hat. Genutzt wird ein maschinelles Lernverfahren, das entlang der Fragestellung im Projekt möglichst viel Erfahrungswissen zu Impactpotenzialen im Quelltyp Projektabschlussbericht öffentlich geförderter Projekte generiert, um maschinell die Wahrscheinlichkeit von Impact auf unterschiedliche gesellschaftliche, wirtschaftliche, technologische, politische, kulturelle usf. Bereiche zu prognostizieren. Das Pilotprojekt hat dies zunächst ausschließlich anhand des Quelltyps Projektabschlussbericht aus der inhaltlich-thematischen Domäne Mobilität versucht. Das hier zusammenfassend dargestellte Hauptprojekt hatte den Kernauftrag, die prinzipielle Funktionalität der Methode *TextTransfer* auf andere Quelltypen und andere Domänen auszudehnen, um seine Funktionalität zu stabilisieren und zu skalieren.

Das Vorhaben folgt hier im Wesentlichen den Erfordernissen eines maschinellen Lernverfahrens (supervised machine learning). Hierfür benötigt das Lernverfahren vorab Informationen, die einen Lerndatensatz für das Methodentraining hinsichtlich des Messgegenstands Impact eindeutig und unterscheidbar klassifizieren. Die Präzision der Methode in der Evaluationsphase unklassifizierter Daten steigt mit der Größe und Heterogenität des Lerndatenkorpus bzw. mit der Güte der in der Trainingsphase vorverabreichten Impactinformationen. Wesentliche Aufgabenstellung in *TextTransfer II* war es demnach, eine hinreichend große Stichprobe als Lern- und Evaluationsdatensätze zu ziehen sowie ein geeignetes Schema und ein stabiles Erhebungsverfahren für Metadaten zur Klassifikation der Datensätze zu erstellen.

Folgende Arbeitsschritte waren daher durchzuführen.

1. Definition der Forschungsfrage, des Messgegenstands und der Grundsätzlichen Methodik (AP1)
2. Auswahl der Quelltypen und der Domänen (AP2).
3. Generierung der Stichprobe (AP 3)
   - 3.1 Überführung der Abgabeformate der Projektendberichte in ein maschinenlesbares Format.
   - 3.2 Erstellung von Kategorienschemata für eine maschinenlesbare Klassifikation von Impactindikatoren.
     - 3.2.1 Top-Down-Ansatz (deduktiver Ansatz): Erhebung von Impactinformationen aus Referenzdaten außerhalb der Datensätze (Zeitungstexte/DeReKo, Webarchive, Projektdokumentation) zur Klassifikation ganzer Berichtstexte und Projekte. Zuordnung textinhärenter Features zu textextern erhobenen Referenzdaten.
     - 3.2.2 Buttom-Up-Ansatz (induktiver Ansatz): Erhebung von Impactinformationen anhand der Identifizierung von Impactindikatoren innerhalb der Datensätze.
   - 3.3 Erstellung der Stichprobe entlang Quelltyp, Domäne und Kategorienschema; Annotation des Lerndatensatzes mit Impactinformationen.
   - 3.4 Definition von Verfahren zur Maßnahmenoptimierung.
     - 3.4.1 Automatisierung der Formatkonvertierung.
     - 3.4.2 Anpassung der Kategorienschemata.
     - 3.4.3 Automatisierung der Referenzdatenerhebung im Top-Down-Ansatz (deduktiv).
     - 3.4.4 Automatisierung der Informationsextraktion im Buttom-Up-Ansatz (induktiv).
4. Durchführung des maschinellen Lernens (AP 4).
   - 4.1 Experimentelle Evaluation verschiedener Finetuning-Ansätze.
   - 4.2 Abgleich der Methodenpräzision nach Finetuning und Erhebungsverfahren von Impactinformationen (deduktiv vs. induktiv).
5. Erarbeitung von Vorschlägen zur Verbesserung der Maschinenlesbarkeit im Abgabeformat von Projektabschlussberichten (AP 5).

FKZ 01IO2002A
Version 1.1

6. Erarbeitung von Vorschlägen für die Methodennutzung (AP 6) und zur Disseminierung von Projektergebnissen (AP 7).

Die Arbeitsschritte im Detail werden im Folgenden dargestellt.

3.1.1. AP1: Bezugsrahmen

AP1 diente der Schaffung von Voraussetzungen, um den Projektgegenstand zu vertiefen und weiterzuentwickeln mit dem Ziel, die Belastbarkeit der Methode zu testen und zu stabilisieren.

Die Forschungsfrage des Vorhabens musste basierend auf den Ergebnissen aus *TextTransfer I* unter Berücksichtigung der erweiterten Arbeitsfelder mit zum Teil – im Vergleich zum Pilotprojekt – neuen Inhalten spezifiziert, dokumentiert und Handlungsfelder abgeleitet werden. Zu diesem Zweck wurden Parameter zur Erweiterung bestehender Datensammlungen und zur Identifikation neuer relevanter Domänen aufgestellt.

Der Untersuchungsgegenstand „Forschungswissen“ wurde in der Pilotphase primär zum Funktionsnachweis anhand von Projektendberichten fixiert. Im Sinne einer umseitigen Stabilisierung der Methode, die mittels des automatisierten Abgleichs von Forschungstexten und klassifizierenden Metainformationen auf die Prognose von Impactwahrscheinlichkeiten zielt, waren im Hauptprojekt in den Arbeitspaketen 2 und 3 alternative Textformate auf ihre Nutzbarkeit für die Methode zu eruieren. In der Folge war dann ein breiteres Arsenal an Referenzdaten für das Maschinelle Lernen zu generieren, anhand dessen unterschiedliche Formen der Rezeption und des Impacts von Forschung abzulesen sein sollte. Hierfür sollten zusätzliche mögliche Quelltypen – neben dem der Projektberichte aus dem Pilotprojekt –, die sich für die Anwendung der Methode eignen, identifiziert bzw. evaluiert werden. Um die Methoden künftig auch auf größeren und mit Blick auf das Spektrum der Domänen heterogenen Datenmengen zu testen, sollte außerdem u.a. eine weitere Domäne für AP 2 Anwendungsfälle – neben den Domänen Mobilität, Künstlichen Intelligenz und Germanistischen Linguistik – aus dem Bereich der Geisteswissenschaften, konkret die der Musikwissenschaften, herangezogen werden. Zudem sollten Möglichkeiten der institutionellen Verankerung bei den Beteiligten sollten aufgezeigt werden. Details zum Thema Auswahl der Domänen und Quelltypen vgl. im Folgenden.

FKZ 01IO2002A
Version 1.1

Um zusätzliche Möglichkeiten der Erhebung weiterer Daten zu generieren und die Datenbasis, mit der die Methode aus *TextTransfer I* weiter optimiert werden konnte, zu erhöhen, wurden im Rahmen von AP 1 außerdem zwei Aufgabenbereiche identifiziert, die sich auf Grund aktueller Entwicklungen bei den Projektpartnern zu Projektbeginn als mögliche Hebel hinsichtlich Ressourcenumverteilung zu erweisen schienen, die aber zum Zeitpunkt der Antragsstellung noch nicht abzusehen waren.

Zwei wesentliche Arbeitsschritte zur Durchführung der Methode leiteten sich unmittelbar aus den Vorarbeiten im Pilotprojekt ab und mussten für die geänderten Rahmenbedingungen im Hauptprojekt aufgegriffen und vertieft werden. Zum einen handelte es sich um den Bereich der **Referenzdatenerhebung**, wie sie in **Fehler! Verweisquelle konnte nicht gefunden werden.** vorgesehen war. Referenzdaten dienen der Klassifizierung des Impacts in Quelltexten für das maschinelle Lernverfahren. In *TextTransfer I* wurden zur Erhebung der Referenzdatenbasis vornehmlich Interviews mit den Autoren des Quelltyps Projektbericht durchgeführt, um klassifizierende Aussagen zum Impact einzelner Projekte zu gewinnen. Der Aufwand dieses Ansatzes entpuppte sich jedoch im Laufe des Pilotprojekts als zu groß für die praktische Umsetzung und Nachnutzung bei potenziellen Anwendern. Dieser Ansatz wurde daher durch die Adaption eines Online-Umfragetools prozessoptimiert.

Eine weitere Prozessanpassung war beim Thema Vor-/Aufbereitung der Daten für das Maschinelle Lernverfahren, ebenfalls Teil von AP 3 Stichprobe, im Bereich der **Konvertierung** vorgesehen. Anstelle der sehr aufwendigen Konvertierung der im PDF-Format vorliegenden Daten in die Markup-Sprache TEI XML (i5) konnte auf die zwischenzeitliche Entwicklung der Automatisierung der Umwandlung von PDFs in Textdaten zurückgegriffen werden, die vom Projektpartner TIB eingesetzt wurde.

Dabei konnten Synergien mit dem parallel laufenden, ebenfalls vom BMBF geförderten Projekt TrenDTF genutzt werden, vgl. https://projects.tib.eu/trendtf/. Der Text aus den Forschungsberichten im PDF-Format wurde zunächst mit Hilfe der Java-Bibliothek PDF Escape extrahiert. Während des Extraktionsprozesses wurde festgestellt, dass einige der PDF-Dokumente gescannt sind, die mit OCR-Techniken in Text umgewandelt wurden. Insgesamt konnten für die Projekte TrenDTF und *TextTransfer II* (Hauptprojekt) 67.600 Dokumente mit beiden Textextraktionsmethoden extrahiert werden, von denen 65.907 PDF-Dokumente mit PDF Escape extrahiert wurden, während 1.413 PDF-Dokumente mit

OCR in Text umgewandelt wurden. 2.765 PDF-Dokumente wurden aufgrund der Speicherbegrenzung der eingesetzten Tools ignoriert.

AP1 war zu Projektende planmäßig abgeschlossen.

3.1.2. AP2: Anwendungsfälle

Der vorliegende Entwicklungsstand der Methode aus *TextTransfer I* gestattete linguistische Analyseverfahren auf einer den Quelltypus Projektbericht und die Domäne Mobilität zu Testzwecken eingegrenzten Datenbasis. Das Arbeitspaket zielte auf die Schaffung einer erweiterten Datengrundgesamtheit für das maschinelle Lernverfahren. Die Funktionalität der Methode sollte hierbei durch Hinzuziehung neuer Quelltypen – und somit neuer Formate – sowie neuer Domänen – und somit thematisch unterschiedlichen Texten – erprobt und erweitert werden.

3.1.2.1. Domänenerweiterung für den Quelltyp Projektberichte

Das Transparenzgebot für die Verwendung steuerlicher Mittel legt öffentlich geförderter Forschung die Verpflichtung auf, ihre Ergebnisse in Projektabschlussberichten darzustellen. Auch hinsichtlich ihrer öffentlichen Verfügbarkeit bei der als Datengeber kooperierenden TIB erschien dieser Texttyp als relativ niedrigschwellig nutzbare Quelle im Vorhaben. Nicht zuletzt aufgrund ihrer im Vergleich zu anderen Formaten wissenschaftlicher Publikation 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 ausgesucht.

Als erweiterter Bezugsrahmen wurden vom Verbundpartner IDS entsprechend Forschungsprojekte ausgewählt,

- die aus öffentlich geförderten Verbundprojekten stammen,
- deren Projektpartner sowohl aus öffentlicher Forschung als auch der freien Wirtschaft in Deutschland kommen,

- deren deutschsprachige Projektabschlussberichte in der TIB frei zugänglich in digitaler Form vorliegen, wobei ein Projekt einen Abschlussbericht stellvertretend für das Gesamtprojekt und/oder Einzelberichte pro Projektpartner auszuweisen hat.

3.1.2.1.1. „Alt"-Domäne Mobilität

Projekte der Domäne Mobilität, die von unterschiedlichen Projektträgern gefördert und alle in der TIB unter dem Recherchebegriff „Mobilität" geführt werden, wurden im Verlauf des Projektes erneut zwecks Abschluss bzw. Vertiefung des Ansatzes aus *TextTransfer I* genauer betrachtet. Die Domäne wurde im Pilotprojekt aufgrund hoher gesellschaftlicher Relevanz und potenziell hohem Impactpotenzial herangezogen. Die Erfahrungen aus *TextTransfer I* hatten die Eignung der Domäne als Evaluierungsgegenstand bestätigt. Eine große Datenbasis, wie sie mit der Domäne Mobilität im Vergleich zu *TextTransfer I* vorliegt, diente außerdem der Erhöhung der Skalierbarkeit der Methode.

3.1.2.1.2. Neu-Domänen

Für den Quelltyp der Projektberichte sollten zwecks Aufbau einer neuen Datenbasis weitere Domänen aus dem Bestand des Projektpartners TIB betrachtet werden, von denen einerseits entlang des bisherigen Suchfilters eine hohe Impactwahrscheinlichkeit zu erwarten war, die aber auch vor allem eine Bedeutung für die Anwendung im Rahmen der Impact- und Transferaktivitäten der beteiligten Partner haben sollten. Zudem sollte im Sinne eines breit aufgestellten Transferverständnisses und eines ganzheitlichen Transferauftrages über alle Disziplinen hinweg eine weitere Domäne aus dem Bereich Geisteswissenschaften herangezogen werden. Entsprechend den Vorgaben wurde folgende weitere Domänen hinzugezogen:

Domäne Künstliche Intelligenz

Die Domäne Künstliche Intelligenz wurde als Thema von hoher gesellschaftlicher Relevanz ausgewählt. Das Thema war zudem Gegenstand des Wissenschaftsjahres 2019[7]. Im Hinblick auf ihr Impact- und Transferpotenzial war sie für die methodische Ausrichtung des Projektpartners IDS von zentraler Be-

[7] https://www.wissenschaftsjahr.de/2019/

deutung: Das Projekt wurde aus der Abteilung Digitale Sprachwissenschaft des IDS geleitet und bediente sich bei der Entwicklung der Methode der Technik des maschinellen Lernverfahrens, einem Teilgebiet der Künstlichen Intelligenz.

Domäne Germanistische Linguistik

Um die Methode auf eine Domäne im Kern-Forschungsinteresse des Projektpartners IDS als sprachwissenschaftliche Einrichtung zu erweitern und gleichzeitig das Thema Transfer- und Impactpotenzial am IDS selbst zu unterstützen, wurde als weitere neue Domäne die der germanistischen Linguistik ausgewählt.

Domäne Musikwissenschaften

Neben den bereits genannten sollte eine weitere Domäne aus dem Bereich der Geisteswissenschaften herangezogen werden. Die Wahl fiel aus zwei wesentlichen Gründen auf die Domäne Musikwissenschaften: Einerseits sollte ein zusätzliches thematisches Feld aus den Bereichen Geistes- und Kulturwissenschaften erschlossen werden, um zu eruieren, welche Ergebnisse ein maschinelles Lernverfahren generiert, das sich bei der Stichprobenanalyse nicht auf klassische Indikatoren und Formate des Transfers stützen kann. Dabei erschien es interessant, den Bestand an geeigneten Referenzdaten zu erweitern. Andererseits sollte sich idealerweise hier auch im Projektverlauf ein weiterer Quelltyp ergeben und dem Vorhaben eine Verifizierbarkeit der Methodenfunktionalitäten ermöglichen. Als Partner sollte das Musikwissenschaftliche Seminar der Universität Detmold/Paderborn, das über Kooperationen im Rahmen der Initiative Nationale Forschungsdateninfrastruktur (NFDI) – namentlich der beiden Verbünde Text+ sowie NFDI4Culture –, mit dem Projektpartner IDS verbunden ist, gewonnen werden. Vertiefte Gespräche und Briefings zum Ansatz des Projektes *TextTransfer* mit dem Musikwissenschaftlichen Seminar der Universität Detmold/Paderborn als auch erste Recherchen zum Thema Anwendungsfälle – speziell der Recherche einer Grundgesamtheit der Domäne Musikwissenschaften als auch nach einem potenziellen neuen Quelltyp – durch das Musikwissenschaftliche Seminar der Universität Detmold/Paderborn ergaben, dass für die Domäne Musikwissenschaften weder im Bestand

des Seminars selbst als auch erweitert bei spezifischen Partnereinrichtungen der Universität Detmold/Paderborn eine nennenswerte Grundgesamtheit des Quelltyps Projektberichte identifiziert werden konnte, die nicht bereits auch im Bestand des Projektpartners TIB nach Durchlauf einer ersten Recherche zum Thema Musikwissenschaften vorhanden gewesen wäre, so dass es zu keiner Kooperation zwischen dem Projekt und der Universität Detmold/Paderborn kam.

Um diese Domäne dennoch als Stichprobe dem Maschinellen Lernverfahren zuzuführen, wurde letztendlich daher erneut auf den Bestand der TIB zurückgegriffen.

Der Projektpartner IDS hatte für die Etablierung der Grundgesamtheiten einen kriteriengestützten Suchfilter je Domäne und innerhalb einer Domäne nach Projekten mit ihren jeweiligen Berichten mit den folgenden Kriterien eingerichtet, die der Projektpartner TIB auf seinen Bestand anwandte. Der Filter diente der Identifizierung geeigneter Quelldatensätze und stützte sich auf folgende Indikatoren: Verbundname Gesamtprojekt, Akronym, Verbundnummer, Name Teilprojekt, Name und Kontaktdaten Forschungseinrichtung/Zuwendungsempfänger, Projektleiter, Fördernummer, Projektträger, Projektlaufzeit, öffentliche Verfügbarkeit des Berichts, digitale Verfügbarkeit des Berichts im TIB-Bestand, Berichtstypus (Konsortial-, Gesamtbericht/Individualbericht), Seitenanzahl, Link zu PDF in der TIB-Bibliothek. Die jeweilige Grundgesamtheit je Domäne wurde in Form einer Excel-Tabelle an das IDS zwecks finaler Sichtung der Zusammenstellung mit Blick auf die finale Eignung eines Datensatzes unter Berücksichtigung der Kriterien des erweiterten Bezugsrahmens für die Stichproben aus AP 3 übergeben.

3.1.2.2. Quelltyperweiterung

Neben dem Quelltyp Projektendberichte sollten weitere Quelltypen hinsichtlich des erweiterten Transferbegriffs als Lerndaten evaluiert werden. Hierzu wurden folgende Ansätze verfolgt:

Ein erster Ansatz, der im Austausch mit dem Projektträger DLR entstand, bezog sich auf die vom Projektträger sekundär erstellten, öffentlich gemachten Projektsteckbriefe, die übersichtsartige Angaben zu Anträgen von Gesamtprojekten und Verwertungsplänen enthalten. Hier sollte überprüft werden, inwieweit dieser Quelltyp dem in *TextTransfer II* gewählten methodischen Ansatz genüge. Die Recher-

che ergab jedoch, dass der Quelltyp Projektskizze nicht öffentlich zur Verfügung steht und die Archivierungspraxis einzelner Projektträger stark voneinander abweicht. Ein neuer Quelltyp ließ sich für die Gattung öffentliche Projektsteckbriefe abschließend nicht identifizieren.

Ein weiterer Ansatz war der der Prüfung von Forschungsanträgen als möglicher zusätzlicher Quelltyp. Dabei sollten Projektanträge zur Stichprobe der Domäne Mobilität aus *TextTransfer I* herangezogen werden, da für diese bereits das Impactpotenzial bekannt ist. Als besondere wissenschaftliche Herausforderung stellten sich hierbei die evolutionäre Distanz von Wissenstand und Kapazität der Projektbeteiligten in der Antragsphase hin zum tatsächlich erzielten Projektergebnis dar. In der Praxis gestaltete sich jedoch der Aufwand der Einholung einer Nutzungsgenehmigung dieses nicht-öffentlichen Quelltyps sowie die Eruierung von rechtsstabilen Ansprechpersonen als noch aufwendiger und sensibler als die Recherche von externen Referenzdaten nach dem deduktiven Ansatz aus dem Evaluierungsprojekt *TextTransfer I*. Eine Reproduktion dieses Vorgehens sowohl innerhalb des vorliegenden Projektplans von *TextTransfer* selbst als auch bei der angestrebten Nachnutzung der Methode im Forschungsalltag erschien daher kapazitär nicht vertretbar. Das Thema wurde somit nicht weiterverfolgt.

Wie bereits im vorangegangenen Kapitel zum Thema Neudomäne Musikwissenschaften beschrieben wurde ein weiterer Ansatz – die Recherche eines weiteren Quelltyps im Zusammenhang mit der Domäne Musikwissenschaften – aufgrund des Nichtzustandekommens einer Kooperationsvereinbarung mit dem Musikwissenschaftlichen Seminar der Universität Detmold/Paderborn nicht weiter vorangetrieben.

Neben den oben genannten Ansätzen wurde außerdem die Idee eruiert, einen neuen Quelltyp mit Hilfe von Informationen aus dem Internet zu generieren: Eine ganze Kategorie möglicher Datenquellen sind dabei öffentlich zugängliche Online-Quellen – darunter z. B. verschiedenartige Websites, Presseerklärungen sowie Social-Media-Präsenzen von Forschungseinrichtungen, aber auch journalistische Angebote im Web. Allein die Erwähnung eines Forschungsprojekts sowie deren Urheber, Häufigkeit, Kontexte und Zeiträume der jeweiligen Erwähnungen könnten wertvolle ergänzende Hinweise im Mosaik von Mustern der Verwertbarkeit offenbaren. Einzelne, aktuelle Quellen dieser Art lassen sich durch Linksammlungen, Websuchmaschinen und Website-interne Suchfunktionen finden und prüfen.

Schwieriger würde es sein, eine große Anzahl ausgewählter Websites mit reproduzierbarer Umfassendheit und Präzision auf Stichwortkombinationen hin zu untersuchen, zumal wenn es gelten würde, umfassende Sammlungen heruntergeladener Web-Dokumente aus der Vergangenheit zu berücksichtigen. In diesem Zusammenhang hätten in Kooperation mit umfassenden Web-Archiven[8] Datenquellen erschlossen werden müssen. Web-Archive erlauben dabei keine Stichwortsuche. Mit der Technik des fokussierten Crawlens wäre es allerdings möglich, bedarfsgerecht zugeschnittene aktuelle oder historische Kollektionen von Web-Dokumenten erstellen zu lassen, die dann lokal auf Rechnern von Mitarbeiterinnen und Mitarbeitern des Projekts analysiert werden könnten. Hierbei wäre die Herausforderung die Bewältigung des Umgangs mit verhältnismäßig großen Datenmengen, d.h. Suchkriterien zu entwickeln, mit denen sich den Kollektionen der Web-Dokumente die erforderlichen Informationen entziehen lassen, sowie die Ergebnisdaten schließlich in einem Modell zu erfassen, das Überblick und Vergleichbarkeit zu gewinnen helfen würde.

Nach erster Prüfung wurde dieser Ansatz einer neuen Quelltyperschließung mittels Daten aus Web-Archiven als interessant und durchführbar befunden. Es wurde dem Projektträger DLR entsprechend vorgeschlagen, angesichts der zu erwartenden hohen Aufwände für diesen lohnenden Ansatz einen systematischen Suchdienst für alternative, Impact-relevante Daten aus dem Web-Archive in seine künftigen strategischen Überlegungen für weiterführende Forschungsprogramme aufzunehmen.

Letztendlich wurde die Entscheidung im Projekt getroffen, die Anstrengungen nach der Identifikation eines weiteren Quelltyps neben dem der Projektendberichte nicht weiter zu verfolgen, sondern sich auf die sich – wie sich nach einer ersten Recherche ergab – weit vielversprechendere erscheinende Suche nach neuen Impact-Referenzdatentypen zu konzentrieren. Details hierzu sind im weiteren Verlauf des Kapitels unter AP 3 zu entnehmen.

AP2 war zu Projektende planmäßig abgeschlossen.

[8] U.a. dem Unternehmen Internet Archive (https://archive.org)

3.1.3. AP3: Stichprobe

Der methodische Ansatz des Verbundes sah die Generierung von geeigneten Stichproben vor, die als Bezugsbasis für das maschinelle Lernen dienen sollten. Hierfür griffen für den Quelltyp Projektberichte die Prozesse der physischen Aufbereitung (Strukturierung und Konvertierung) durch den Projektpartner TIB sowie der informationellen (u.a. Auswahl, Referenzdatenerhebung, Kategorisierung, Annotation) Datenauf- und Vorbereitung durch den Projektpartner IDS ineinander.

Mit den in AP 2 Anwendungsfälle am Vorgehen im Pilotprojekt orientierten Eingrenzungen konnten für die genannten Domänen folgende Stichproben aus dem TIB-Bestand für die weitere Bearbeitung extrahiert werden:

- Domäne Mobilität (Mob): 239 Projekte mit 906 Einzelberichten[9]
- Domäne Künstliche Intelligenz (KI): 116 Projekte mit 497 Einzelberichten
- Domäne Linguistik (Ling): 40 Projekte mit insgesamt 89
- Domäne Musikwissenschaften (MuWi): 25 Projekte mit insgesamt 58 Einzelberichten

3.1.3.1. Konvertierung

Sobald die Stichproben definiert waren, sah der weitere Prozess die Vor- und Aufbereitung der zugrundeliegenden Daten für das maschinelle Lernen. Bei der Quellenbasis der TIB für den Quelltyp Projektberichte handelte es sich um PDF-Dokumente, die für die weitere Bearbeitung in ein maschinenlesbares Format, u.a. in das txt-Datenformat, gebracht werden mussten. Der Projektpartner TIB war, in Abweichung zur Antragsstellung, auf Grund aktueller Entwicklungen zu Projektbeginn von *TextTransfer II* in der Lage, PDFs automatisiert in das txt-Datenformat umzuwandeln. Anstelle der sehr aufwendigen Konvertierung der im PDF-Format vorliegenden Daten in die Markup-Sprache TEI XML (i5) konnte auf die zwischenzeitliche Entwicklung der Automatisierung der Umwandlung von PDFs in Textdaten zurückgegriffen werden, die vom Projektpartner TIB eingesetzt wurde. Dabei konnten Synergien mit dem parallellaufenden, ebenfalls vom BMBF geförderten Projekt TrenDTF genutzt werden (vgl. https://projects.tib.eu/trendtf/). Der Text aus den Forschungsberichten im PDF-Format wurde zunächst mit Hilfe

[9] Bestehend aus 148 Projekten mit insgesamt 515 Einzelberichten aus TextTransfer II, zzgl. der Stichprobe aus *TextTransfer I* mit 91 Projekten mit insgesamt 391 Einzelprojekten.

der Java-Bibliothek PDF Escape extrahiert. Während des Extraktionsprozesses wurde festgestellt, dass einige der PDF-Dokumente gescannt sind, die mit OCR-Techniken in Text umgewandelt wurden. Insgesamt konnten für die Projekte TrenDTF und *TextTransfer II* (Hauptprojekt) 67.600 Dokumente mit beiden Textextraktionsmethoden extrahiert werden, von denen 65.907 PDF-Dokumente mit PDF Escape extrahiert wurden, während 1.413 PDF-Dokumente mit OCR in Text umgewandelt wurden. 2.765 PDF-Dokumente wurden aufgrund der Speicherbegrenzung der eingesetzten Tools ignoriert.

Eine Prozessoptimierung im Bereich der Konvertierung konnte erfolgen, indem die Daten anhand eines eigens dafür vom IDS entwickelten Workflows für die weitere Bearbeitung im Rahmen von **Fehler! Verweisquelle konnte nicht gefunden werden.** durch den Projektpartner IDS aufbereitet werden konnten. Der für *TextTransfer II* herangezogene Konvertierungsprozess wurde dabei konsequent aus den Erfahrungen des Pilotprojekts entwickelt.[10]

3.1.3.2. Referenzdatenerhebung – Entwicklung eines Online-Umfragetools

Maschinelle Lernverfahren basieren auf künstlichem Erfahrungswissen, das durch die Klassifizierung von Lerndaten durch diskriminierende Kategorien gesteuert wird. Repetitives Lernen vorklassifizierter Daten erzeugt die Fähigkeit automatisierter Identifikation bestimmter diskriminierender Features in Texten (supervised learning). Die Kategorien werden hierzu aus text-externen (deduktiv, top-down) oder -internen (induktiv, bottom-up) Metainformationen etabliert; Aussagen über der Maschine unbekannte, nicht klassifizierter Texte können anhand dieser Referenzen evaluiert werden.[11] Die Verfahren, die in *TextTransfer I* zur Erhebung der Referenzdatenbasis notwendig waren, bedienten sich überwiegend analoger Ansätze (Interviews, manuelle Extraktion und Annotation von impactrelevanten Textpassagen) und waren entsprechend ressourcen-intensiv hinsichtlich Personal bzw. Zeit. Mit Blick auf die angestrebte institutionelle Nutzungsroutine der Methode waren vergleichbare Ressourceninvestitionen in die Vorbereitung eines domänen- und quelltypangepassten, maschinellen Lernens nicht realistisch. Für das laufende Projekt wurde daher definiert, insbesondere die analogen Interviews zur

[10] Vgl. hierzu das Pilotprojekt *TextTransfer I* unter https://www.tib.eu/de/suchen/id/TIBKAT:1747197327/TextTransfer-Pilot-korpusgestützte-Erkennung-von?cHash=f56c7df1117392f268358ce611858ca4

[11] Details hierzu sind dem Abschlussbericht von TextTransfer I/Kap. 3 zu entnehmen, der frei zugänglich unter https://www.tib.eu/de/suchen/id/TIBKAT:1747197327/TextTransfer-Pilot-korpusgestützte-Erkennung-von?cHash=f56c7df1117392f268358ce611858ca4 abrufbar ist.

Informationsbeschaffung des Impact- und Transferpotenzials im Rahmen des Top-Down-Ansatzes anzupassen, um stattdessen auf eine Adaption des am IDS entwickelten Online-Umfrage-Tools zurückzugreifen:

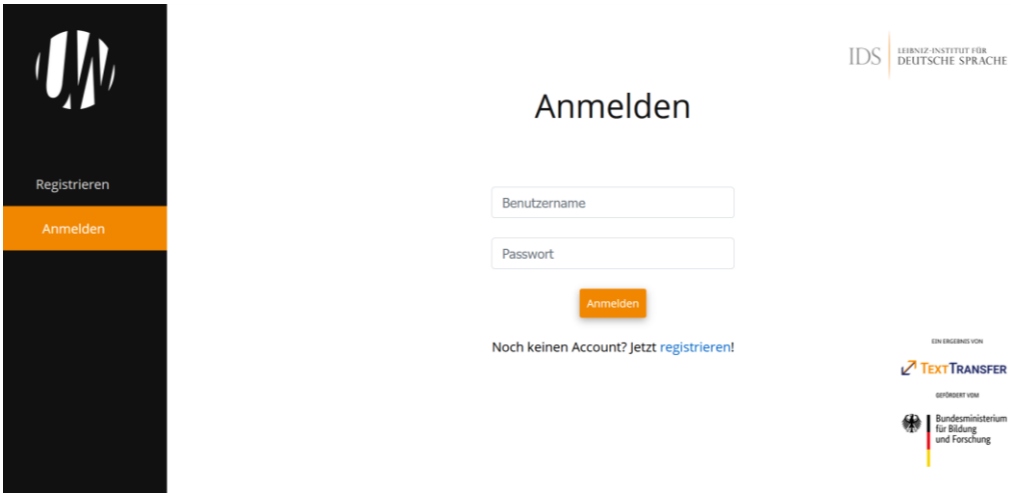

Abbildung 3-1 Für die Nutzung des Online-Umfrage-Tools „umfragewissen.texttransfer.org" ist eine Registrierung erforderlich.

Das Online-Umfrage-Tool mit dem Namen „umfragewissen.texttransfer.org“[12] umfasste folgende Grundfunktionalitäten:

- die Erstellung, Überarbeitung und Löschung von Umfrageformularen,
- eine Live-Vorschau von Umfrageformularen,
- die Einstellungsmöglichkeit einer individuellen Teilnahmebestätigungsnachricht für Umfrageteilnehmende,
- die automatische Generierung von Links zur Einladung potenzieller Probandinnen und Probanden,
- die Download-Möglichkeit von Umfrageergebnissen als CSV-Datei,

[12] https://www.umfragewissen.texttransfer.org/

- eine Tracking-Funktion zur Detailansicht einzelner Antworten sowie zur sekundengenauen Zeiterfassung einzelner Teilnehmerinnen und Teilnehmer.

Insbesondere dem letztgenannten Punkt, der Tracking-Funktion, kam eine besondere Bedeutung zu. Diese Zeiterfassung ermöglichte es, Umfragen bereits vor der Veröffentlichung umfassend im Hinblick auf die Teilnahmedauer zu testen sowie die tatsächliche Bearbeitungszeit eines Teilnehmenden während einer Befragung zu ermitteln. Dies hatte den Vorteil, den Teilnehmenden bereits vor Beginn der Umfrage einen relativ genauen Zeitrahmen bezüglich der Bearbeitung geben und damit mögliche zeitliche Barrieren im Vorfeld abklären und potenzielle Abbruchquoten seitens der Teilnehmenden während einer Befragung möglichst reduzieren zu können. Die effiziente Auf- und Vorbereitung der Umfragen konnte somit nicht nur zur Vermeidung einer hohen Abbruchquote, sondern auch erheblich zu einer guten Rücklaufquote beitragen, was wiederum eine höhere Anzahl zu verwertbaren Aussagen unterstützte.

Das Tool bot sowohl eine Übersicht vergangener Umfragen als auch insbesondere eine schnelle und unkomplizierte Anlage einer Umfrage, ohne eine große Einarbeitung vorauszusetzen.

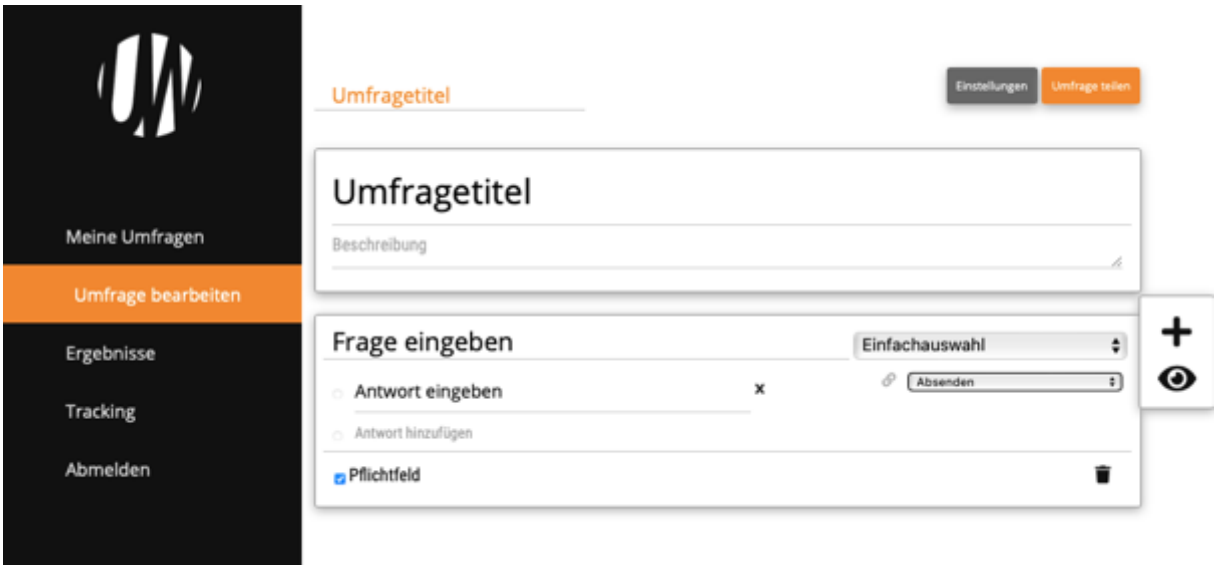

Abbildung 3-2 Erstellung bzw. Überarbeitung einer Umfrage mittels „umfragewissen.texttransfer.org".

FKZ 01IO2002A
Version 1.1

Das Herunterladen der Umfrageergebnisse als CSV-Datei erlaubte außerdem schnelle Auswertungen in diversen Anwendungen.

Wie im Vorangegangenen erwähnt sollte die Nutzbarkeit des Tools mittels einer institutsinternen, anonymisierten Umfrage am IDS getestet werden. Aufgrund der inhaltlichen Nähe der Kompetenz des Partners bot es sich an, die Stichprobe der Domäne Germanistische Linguistik zu verwenden.

Zur Bereitstellung des Online-Umfrage-Tools wurde auf einen IDS-externen Webserver – linode.com[13] – zurückgegriffen. An der Erstellung als auch v.a. an der Auswertung der Ergebnisse war der Unterauftragnehmer G&K beteiligt. Die Konstellation der Nutzung eines externen Servers als auch die Weitergabe von Daten an Dritte erforderten nach Vorgaben der Datenschutzbestimmungen des IDS einen Vertrag zur Auftragsdatenverarbeitung als auch die Anfertigung eines Verfahrensverzeichnisses.[14]

Das Front-End des Tools wurde mit Hilfe von HTML, CSS und JavaScript sowie dem CSS-Framework Bootstrap[15] entwickelt. Das Back-End wurde in Python geschrieben bzw. unter Verwendung des Python-Frameworks Django[16] implementiert. Als Datenbank wurde SQlite eingesetzt.[17] Diese ist standardmäßig in Django integriert und – im Hinblick auf die Performance – für kleinere Webprojekte bereits ausreichend. Im Vergleich zu anderen relationalen Datenbanken, wie MySQL oder PostgreSQL, war sie zudem einfacher einzurichten und insbesondere für prototypische Zwecke bestens geeignet.[18]

Die Entscheidung zur Nutzung eines projekteigenen Umfragetools fiel insbesondere mit Blick auf die Nachhaltigkeit des Forschungsprojekts: Der Vorteil im Vergleich zur Nutzung eines externen Tools, wie z. B. Google Forms o. ä., lag neben der Vermeidung von möglichen Komplikationen hinsichtlich datenschutzrechtlicher Belange vor allem in der Möglichkeit, individuelle Funktionen, welche für künftige Methoden der Impacterfassung evtl. notwendig sein könnten, zu entwickeln und zu integrieren. Dies

[13] https://www.linode.com/
[14] Die entsprechenden Verträge und Unterlagen sind am IDS gespeichert und können bei Bedarf per Email an bopp@ids-mannheim.de angefordert werden.
[15] https://getbootstrap.com/
[16] https://www.djangoproject.com/
[17] https://www.sqlite.org/index.html
[18] Aufgrund des Umfangs der Skripte, sind diese nicht Teil des Berichts, können jedoch auf Anfrage zur Verfügung gestellt werden.

wäre bei Verwendung eines externen Tools, welches in der Regel keinen bzw. nur eingeschränkten Zugriff auf den zugrundeliegenden Quellcode erlaubt, nicht realisierbar.

3.1.3.3. Kategorienschema

Für *TextTransfer II* galt es durch den Projektpartner IDS zu überprüfen, inwieweit sich das bisherige, im Evaluierungsprojekt *TextTransfer I* genutzte, mit Blick auf die Domäne Mobilität entwickelte Kategorienschema zur Impacterfassung der untersuchten Forschungsprojekte, auch für die Erweiterung der Domänen – konkret für die Domänen Künstliche Intelligenz (KI), Linguistik (Ling) und Musikwissenschaften (MuWi) – eignete.

Im Pilotprojekt *TextTransfer I* wurde mit folgenden Überlegungen gearbeitet:

> *„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 (…) her gemessen wird, kann sich indes in sehr unterschiedlichen Formen zeigen und wurde für die Messung entsprechend (…) 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. (…) 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.“*[19]

[19] Vgl. *TextTransfer* I/Kap. 3/AP 2 unter https://www.tib.eu/de/suchen/id/TIBKAT:1747197327/TextTransfer-Pilot-korpusgestützte-Erkennung-von?cHash=f56c7df1117392f268358ce611858ca4

Für die Erfassung des tatsächlichen Impacts der untersuchten Projekte der Domäne Mobilität wurde dabei die Definition von „Impact" als „durch Transfer/Verwertung angestoßene nachweisbare Effekte einer existenziellen, gesellschaftlichen, wirtschaftlichen oder ökologischen Veränderung" zugrundegelegt.

Die Definition des Impactbegriffes war ein zentraler Baustein zur Impacterfassung. Eine qualitative Aufbereitung der Recherche- und Interviewstudie für tatsächlich erfolgte Verwertung aus dem Pilotprojekt *TextTransfer I* ergab, dass die vorliegende Definition von „Impact" aus dem Pilotprojekt auf Grund der Domänenerweiterung nur bedingt für eine Erhebung des tatsächlichen Impacts für *TextTransfer II* geeignet war. In der Formulierung „durch Transfer/Verwertung angestoßene nachweisbare Effekte einer existenziellen, […] Veränderung" wurde ein notwendiger kausaler Zusammenhang zwischen „Transfer" und „Effekt" konstatiert, der im Rahmen der recherchegestützten Erhebung in *TextTransfer I* nur vereinzelt nachgewiesen werden konnte. Schließlich war ein Nachweis darüber, dass eine Veränderung in einem gesellschaftlichen Bereich auf den Transfer von Forschungsergebnissen zurückzuführen ist, meist nur mit erheblichem Ressourceneinsatz zu erbringen. Darüber hinaus wurde in der Definition die Unterscheidung von „[...] existenziellen, gesellschaftlichen, wirtschaftlichen oder ökologischen Veränderungen" gemacht, die sich für die Impacterfassung der Domäne Mobilität als geeignet erwies. Die Frage stellte sich, inwieweit sich durch diese Definition auch Veränderungen bezüglich der neu hinzugekommenen Domänen Künstliche Intelligenz, Linguistik und Musikwissenschaften abbilden ließen, also mittels einer Definition, die Bereiche wie Kultur, Politik und Recht nicht dezidiert als Impact erfasste und somit möglicherweise den neuen Domänen bei der Impacterfassung nicht gerecht würde. Zwecks Anwendung der Methode unter Berücksichtigung der Domänenerweiterung wurde in *TextTransfer II* daher eine umfassende Literaturrecherche durch den Projektpartner G&K zu gängigen Definitionen von „Impact" sowie Möglichkeiten zur Operationalisierung des Begriffs durch ein anwendbares Kategorienschema durchgeführt. In Anlehnung an anerkannte Definitionen von „**Impact**"[20] konnte dabei folgende **Definition für *TextTransfer II*** erarbeitet werden: „**Eine Auswirkung auf,**

[20] Vgl. z. B. die Definition von „Broader Impacts" der National Science Foundation (USA): https://www.nsf.gov/pubs/policy-docs/pappg20_1/pappg_3.jsp#IIIA2b; Die Definition von „Impact" des Research Excellence Framework (UK): https://www.ref.ac.uk/; Die Definition von "Research Impact" des Australian Research Council: https://www.arc.gov.au/policies-strategies/strategy/research-impact-

eine Veränderung oder ein Nutzen für Wirtschaft, Gesellschaft, Kultur, Politik oder Recht, Technologie oder Umwelt, über den akademischen Bereich hinaus". Diese Definition wurde in ähnlicher Form bereits z. B. im Rahmen des Research Excellence Framework (UK) erfolgreich zur Erfassung von Impact operationalisiert. Sie ermöglichte die Erfassung von Auswirkungen, Veränderungen oder Nutzen eines Forschungsergebnisses über den akademischen Bereich hinaus plausibilitätsgestützt, d.h. ohne die notwendige Erbringung eines Nachweises. Außerdem ermöglichte sie explizit auch die Erfassung von Impact in weiteren gesellschaftlichen Bereichen (z. B. Politik, Kultur) und nannte so auch die Bereiche, die in *TextTransfer I* bereits implizit, über die Operationalisierungsebene, mitgedacht wurden.

Auf Grundlage dieser neuen Definition erfolgte eine Anpassung des Kategorienschemas zur Erfassung von Impact, welches als Operationalisierung der genannten Definition einerseits für die Recherche bzw. Abfrage des tatsächlich erfolgten Impacts als auch für das maschinelle Lernen geeignet sein sollte. Bei der Recherche eines Kategorienschemas für *TextTransfer II,* d.h. unter Berücksichtigung der zusätzlichen Domänen Künstliche Intelligenz, Linguistik und Musikwissenschaften, konnten in einem ersten Schritt die PESTEL-Kategorien als mögliches neues Kategorienschema zur Erfassung von Impact identifiziert werden[21]. Zwar wurde dieses Messinstrument primär für die Bedarfe der Ökonomie entwickelt, doch schien es auch für den im Projekt gewählten Ansatz geeignet zu sein, zumal dieser mit dem Messgegenstand Impact in der Hauptsache die Wirkungen angewandter Forschung und nicht die Urheber des genutzten Wissens selbst betrachtet.

Um die Eignung der PESTEL-Kategorien zur Erfassung des tatsächlichen Impacts spezifisch für die Domäne Germanistische Linguistik zu prüfen, diente außerdem wie bereits erwähnt – die Durchführung einer anonymisierten Kurzstudie am IDS, wobei mithilfe des Umfragetools „umfragewissen.texttransfer.org" insbesondere das (intuitive) Verständnis der PESTEL-Kategorien sowie deren Anwendbarkeit und Vollständigkeit bzw. Erweiterbarkeit überprüft werden sollte. Ein weiterer wichtiger Punkt, den es

principles-framework; Die Definition von „Impact" der Europäischen Kommission: ttps://ec.europa.eu/newsroom/horizon2020/redirection/document/10927)

[21] Vgl. z. B.: Schomaker, Rahel M.; Sitter, Alexander (2020): Die PESTEL-Analyse – Status quo und innovative Anpassungen. Der Betriebswirt 61(1): 9-27. Weblink: https://www.researchgate.net/profile/Rahel-Schomaker/publication/341876583_Die_PESTEL-Analyse_-_Status_quo_und_innovative_Anpassungen/links/5fd6285192851c13fe816690/Die-PESTEL-Analyse-Status-quo-und-innovative-Anpassungen.pdf.

zu evaluieren galt, war die Akzeptanz bzw. Skepsis der Teilnehmenden gegenüber der Impacterfassung auf Basis von PESTEL.

Mittels der Umfrage anhand des Online-Tools konnte, neben den Punkten

- Der Top-Down-Ansatz, der im Pilotprojekt *TextTransfer I* mittels sehr aufwendigem Analogansatz realisiert wurde, kann mittels des Online-Tools extrem zeit- und ressourceneffizient repräsentiert werden.
- Der Ansatz stößt auf eine hohe Akzeptanz bei den Nutzern.

insbesondere Folgendes, die Entwicklung des Kategorienschemas betreffend – ausgehend von der beispielhaften Anwendung beim Projektpartner IDS – festgehalten werden:

- Die PESTEL-Kategorien, die es bzgl. der Erfassung des tatsächlichen Impacts zu überprüfen galt, wurden von den Teilnehmenden positiv bestätigt.[22]

Anspruch an das Kategorienschema war es, hinreichend jene Impactbereiche abzubilden, die im Rahmen von *TextTransfer II* erfasst werden sollten. Hierfür wurde zunächst ein Subset von 12 Berichten (je drei Berichte pro Domäne) in einem Bottom-Up-Ansatz[23] von drei Annotierenden mit frei zu wählenden Impactkategorien annotiert. Diese wiederum sollten – aus subjektiver Sicht – Rückschlüsse auf das Transfer- und Impactpotenzial eines Projektes geben und damit als Grundlage für ein distinktes, domänenübergreifendes Kategorienschema nutzbar sein. Die Texte des Subsets wurden mithilfe eines Topic-Modeling-Ansatzes – ein statistisches Suchschema auf Wahrscheinlichkeitsbasis, das große Datenmengen auf häufige Wortkorrelationen hin exploriert – ausgewählt, um die Diversität der Texte sicherzustellen. Da die Berichte alle in deutscher Sprache gehalten waren, wurden die ausgewählten Stichproben außerdem – um verallgemeinerbare Kategorien für Nicht-Muttersprachler zu definieren

[22] Die Ergebnisse aus der Umfrage (vgl. Anhang) wurden auch bei der Analyse der Messung von Impact mittels medialer Berichterstattungen genutzt vgl. hierzu im weiteren Verlauf des Kapitels zu AP 3 in dem Unterkapitel *Identifikation externer Impact-Referenzen*

[23] Vgl. Abschlussbericht zu *TextTransfer I* https://www.tib.eu/de/suchen/id/TIBKAT:1747197327/TextTransfer-Pilot-korpusgestützte-Erkennung-von?cHash=f56c7df1117392f268358ce611858ca4

– ins Englische übertragen und drei weitere Annotierenden (Englischsprachige) des Unterauftragnehmers UIUC vorgelegt, um das gleiche Verfahren durchzuführen. Den Annotierenden lagen bei diesem Ansatz zu keinem Zeitpunkt irgendwelche Informationen zu tatsächlich nachgewiesenen Transfer- und Impactpotenzialen eines Projektes vor, so dass letztendlich eine relativ objektive Annotation auf Grund des textbasierten Informationsgehalts eines Projektes erfolgen konnte.

Die auf die obengenannte Weise erarbeiteten Kategorien wurden anschließend systematisch miteinander abgeglichen und hinsichtlich ihrer Kompatibilität mit dem Schema nach PESTEL überprüft.

Als wichtige Ergebnisse dieses Vorbereitungsprozesses ließ sich festhalten, dass

- viele der vorgeschlagenen Wortverbindungen mit den Kategorien von PESTEL beschrieben werden konnten,
- der freie Annotationsprozess eine domänenübergreifende Anwendung der Impact-Kategorien vermuten ließ, d.h. das vorgeschlagene Schema für alle vorliegenden Stichproben der Domänen Mob, KI, Ling und MuWi genutzt werden könnte.

Auf der Grundlage der vorgeschlagenen Kategorien und des Abgleichs mit den bestehenden PESTEL-Kategorien wurde anschließend in mehreren Schritten (Annotation, Evaluation, Neuanpassung) ein Codebook mit sieben Hauptkategorien und verschiedenen Unterkategorien entwickelt, das der systematischen Annotation zum Transfer- und Impactpotenzial der Projekte auf Textebene diente[24]. Die Adaption umfasste dabei sowohl das Hinzufügen neuer Kategorien (Ethischer Impact, Impact auf die Umwelt, akademischer Impact sowie die Rest-Kategorie Sonstiger Impact) als auch die Zusammenfassung der in unserem Datensatz infrequenten Kategorien politischer und rechtlicher Impact.

Das Schema umfasste folgende Hauptkategorien:

- Gesellschaftlicher Impact
- Politischer und rechtlicher Impact

[24] Das Codebook in deutscher Sprache: vgl. Anlage; das Codebook in englischer Sprache ist am IDS gespeichert und kann bei Bedarf per Email an bopp@ids-mannheim.de angefordert werden.

- Ethischer Impact
- Ökonomischer Impact
- Ökologischer Impact
- Technischer Impact
- Akademischer Impact

Ein weiterer wichtiger Schritt bestand in der Identifikation von Subkategorien für die sieben Hauptkategorien:

Societal impact: The effects of a project on societal events, groups, or institutions	
Culture / Events	Organization of (cultural) events, workshops, performances, etc.
Education	Education or training, e.g., new learning methods for schools
Physical Health	e.g., alleviation or reduction of diseases, vaccination campaigns, etc.
Life Quality	Quality of life/mental health, e.g., reduced depression, work-life balance
Safety	Traffic safety, hazard prevention, etc.
Mobility	Freedom of movement and mobility
Political and Legal Impact: Utilization of project results in political or legislative contexts	
Regulations	Political regulations and deregulations
Laws	Development/contributions/amendments to laws/legal regulations
Ethical Impact: Ethical effects of a project e.g., raising equality or awareness, charity	
Awareness	Establishment/increase of awareness, perception, or attitudes
Justice	Establishment/improvement of justice and equality
Data Policy	Data privacy and privacy, Open Access
Economic Impact: Utilization of project results for economic/financial developments	
Business Models	Development of business models or other economic strategies
Income	Generation/increase of income, new positions (outside university)
Employee Satisfaction	Employee satisfaction/service quality
Optimizing Processes	Optimization of processes in the economic/business sector
Environmental Impact: Effects of a project on ecological or environmental aspects	
Climate Protection	Environmental/climate/conservation protection
Sustainability	Sustainability of products, methods, etc.
Technical impact: Technologies/models/data developed in the project	
Protoytype	Prototype development: Development of software prototypes
Model Development	Development of models/algorithms/methods
IT Security	IT Security: Establishment/improvement of IT security
Documentation	Documentation: Technical documentation
Data Collection / Release	Data collection/release: Data collection/data release/data publication
Other: Types of non-academic impact that do not fit into one of the other main categories	
Product Development	Development or improvement of a product
PR / Visibility	Promotion of a project; methods to increase the visibility of the project
Knowledge Acquisition	Acquisition/communication of knowledge through the project
Knowledge Transfer	Transfer of knowledge obtained in the project to other areas
Collaborations	Establishment/improvement of collaborations beyond academia
Academic Impact: Effects within the academic domain	
Income Academia	Income of research institutions, funding for hiring new team members
Research Methods	Development/application of new research methods
Learning and Teaching	New/improved learning and teaching methods within academia
Publications	Publication of research results (e.g., in journals/books, at conferences)
Academic Events	Organization of academic events (e.g., workshops, conferences)
Collaborations	Establishment/improvement of collaborations/networks within academia
Future Research	Opening up new perspectives for future research projects
Knowledge Acquisition	Acquisition/conveyance of academic knowledge through the project

Abbildung 3-3 Kategorienschema zur Erfassung des Impacts von Forschungsprojekten in Anlehnung an die PESTEL-Kategorien. Eigene Darstellung.

Um nutzbare Ergebnisse für das Maschinelle Lernen zu generieren, ist zum einen die Quantität, jedoch auch die Qualität der Datensets entscheidend. Für die verwendeten Projektabschlussberichte von Projekten, die immer aus mehreren Partnern bestanden, galt es daher, möglichst präzise Berichte bzw. Berichtsteile auszuwählen, die verwertungs- und impactrelevante Informationen lieferten – unter Ver-

meidung großer Redundanzen. Hier boten sich für den weiteren Annotationsprozess jeweils die Gesamtberichte der Projekte an, da diese übergreifend alle Ergebnisse der einzelnen Projektpartner beinhalten. Die Auswahl sowie die Vor- und Aufbereitung der Stichproben lag dabei – wie bereits im Evaluierungsprojekt *TextTransfer I* – beim Projektpartner IDS.

Die Passgenauigkeit des neu entwickelten Kategorienschemas für alle vier der genannten Domänen, auch in Bezug auf das maschinelle Lernen, zu evaluieren, wurde durch iterative Pilotstudien und Diskussionen des Projektteams im Rahmen von Datensitzungen im Projektverlauf eingehend geprüft – und als positiv beschieden.

3.1.3.4. Datenauf- und Vorbereitung für das Maschinelle Lernen

3.1.3.4.1. Extraktion impactrelevanter Passagen

Im Zuge der Auf- und Vorbereitung der Daten für den maschinellen Lernprozess musste die Stichprobe mit einschlägigen Impact-Informationen durch das IDS auf der Textebene annotiert, also in den Texten manuell kodiert, und anschließend analysiert werden. Da die manuelle Annotation der gesamten Projektberichte, die teilweise mehr als 150 Seiten Umfang hatten, zu zeit- und kostspielig gewesen wäre und um die relevanten, redundanzfreien Textmengen weiter einzuschränken, entwickelte das IDS mit Unterstützung des Unterauftragnehmers UIUC ein Modell zur automatischen Extraktion impact-relevanter Passagen aus allen Gesamtberichten. Dabei handelte es sich um ein Supervised Learning Modell[25], das auf dem Random Forest Algorithmus beruht und pro Satz als Input binär klassifiziert, ob in diesem Satz Impact zum Ausdruck kommt oder nicht. Der Random Forest Algorithmus ist ein Ensemble-Lernverfahren, das zur Trainingszeit eine Vielzahl von Entscheidungsbäumen konstruiert und die Klasse ausgibt, die von den meisten Bäumen ausgewählt wird. Es war für den hier vorliegenden Datensatz in besonderer Weise geeignet, da dieses Verfahren nicht zum Over-Fitting[26] neigt und daher auch für kleine Trainings-Datensätze anwendbar ist.

[25] Maschinelle Lernverfahren basieren auf künstlichem Erfahrungswissen, das durch die Klassifizierung von Lerndaten durch diskriminierende Kategorien gesteuert wird. Repetitives Lernen vorklassifizierter Daten erzeugt die Fähigkeit automatisierter Identifikation bestimmter diskriminierender Features in Texten.

[26] Overfitting bezeichnet im maschinellen Lernen die Überanpassung eines Modells an seine Trainingsdaten. Dies führt in der Regel zu schlechten Ergebnissen auf ungesehenen Testdaten.

Um die Berichte in ein für maschinelle Lernverfahren optimiertes txt-Format zu bringen, wurden diese zunächst mithilfe des Python-Packages "Textract" konvertiert. Anschließend wurden die konvertierten Texte auf der Grundlage textstrukturierender Elemente wie Aufzählungszeichen, Leerzeilen oder Zeilenumbrüche in einzelne Textpassagen aufgesplittet.[27]

Um anschließend Trainingsdaten für das Random Forest Modell zu erzeugen, wurden in einem zweistufigen Annotationsverfahren in 15 Projektberichten alle impactrelevanten Passagen von drei deutschsprachigen wissenschaftlichen Mitarbeitenden des Projektpartners IDS, die über einschlägige Erfahrungen im Bereich Annotation bzw. Computerlinguistik verfügten, gründlich gelesen und die Textstellen (Sätze oder Abschnitte) manuell markiert. Optimiert wurde das Modell durch eine semi-automatisiert erstellte Liste von impact-anzeigenden Wörtern. Hierfür wurde ausgehend von einem Seed Set mit Hilfe der Kookkurrenz-Datenbank CCDB[28] impactindizierende Wörter ermittelt, indem nach Wörtern mit einem ähnlichen Kookkurrenzprofil wie die Wörter „Einfluss" und „Auswirkung" sowie nach Wörtern, die oft im Zusammenhang mit „Einfluss" und „Auswirkung" vorkommen, gesucht wurde. Ergänzt wurde die Liste um die Wörter, die beim manuellen Markieren impactrelevanter Passagen (s.o.) entdeckt wurden. Die Wörterliste wurde anschließend gesichtet und manuell in drei Kategorien unterteilt:

- Kategorie 1 (sehr relevante Impactindikatoren):

  auswirken, Auswirkung, beeinflussen, beeinflußen, Effekt, effektiv, Einfluss, Einfluß, Fortschritt, Impact, nachhaltig, nutzbar, Nutzbarmachung, Potential, Potenzial, umsetzen, Umsetzung, verändern, Veränderung, verbessern, Verbesserung, Verwertung, Verwertungsmöglichkeiten, wirksam, Wirksamkeit, Wirkung

- Kategorie 2 (relevante Impactindikatoren):

[27] Die verwendeten Skripte sind am IDS gespeichert und können bei Bedarf per Email an bopp@ids-mannheim.de angefordert werden.
[28] http://corpora.ids-mannheim.de/ccdb/

beachtlich, Beitrag, beitragen, direkt, Einflussnahme, Einflußnahme, Einflußmöglichkeit, Einflussmöglichkeit, Einsatzmöglichkeiten, hochrelevant, Innovation, innovativ, realisierbar, realisieren, Realisierung, Ziel, zielführend

- Kategorie 3 (möglicherweise relevante Impactindikatoren):

  abschätzbar, abschätzen, anwenden, Anwendung, Anwendungsfall, Anwendungsframework, Anwendungsszenario, Attraktivität, effizient, Entwicklung, Erfolg, Erfolgsaussichten, Ergebnisse, ermöglichen, erreichen, erzielen, Feedback, Frontend, Gewinn, gewinnen, gewinnorientiert, Hauptanwendungsfälle, indirekt, Infrastruktur, infrastrukturell, langfristig, lösen, Lösung, maßgeblich, messbar, meßbar, negativ, neu, nutzen, positiv, produktiv, Projektziele, reagieren, Reaktion, real-world, spürbar, strukturell, Überwindung, unmittelbar, Use Case, Weiterentwicklung, Wertschöpfung, Wettbewerb, Wettbewerbsanalyse, Zukunft, zukünftig, Zweck

Für die Optimierung des Random Forest Algorithmus erwies sich die Kategorie 1 als effektivstes Feature. Hierbei wurde die Anzahl der Vorkommen jedes Signalworts aus Kategorie 1 sowie die aggregierte Anzahl jedes Satzes von Keywords berechnet. Das Random Forest-Modell wurde anschließend mithilfe des sklearn-Python-Packages[29] implementiert. Folgende Parameter erwiesen sich hierbei am effizientesten:

- n_estimators = 1000
- class_weight = 'balanced'
- min_samples_leaf = 6
- max_depth=6

Mit dem Modell konnte eine gewichtete Präzision von 0.74, ein gewichteter Recall von 0.67 und ein F1-Score von 0.69 erzielt werden. Das Modell erzielte bis zu 81 % Accuracy, war übertragbar auf wissenschaftliche Forschungsberichte aus verschiedenen Bereichen – konkret der Künstlichen Intelligenz,

[29] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

der Mobilität, der Germanistischen Linguistik und der Musikwissenschaften – und funktionierte für verschiedene Sprachen (deutsch und englisch).

Weitere Modelle, die erprobt wurden, jedoch zu niedrigeren Performanzen führten, waren die Pre-Trained-Transformer-Language-Modelle BERT[30] und RoBERTa.[31] Da diese Art von Deep-Learning-Modellen jedoch eine sehr große Menge an Trainingsdaten erfordern, waren sie für das hier vorliegende Datenset nicht anwendbar. Die Experimente mit embedding similarities lieferten ebenfalls keine zufriedenstellenden Ergebnisse. Hier wurde die Methode word2vec angewendet, in der Wörter durch Vektoren repräsentiert werden. Anschließend wurde die Kosinus-Ähnlichkeit zwischen den Schlüsselwörtern aus Kategorie 1 (s.o.) und den Inhaltswörtern aus den extrahierten Passagen und interpretierten Passagen mit den höchsten Ähnlichkeitswerten als impact-relevant verglichen. Ein Abgleich mit den annotierten Daten zeigte jedoch auch hier, dass diese Methode für das hier vorliegende Daten-Set und der Aufgabe nicht zum Erfolg führte.

Das Random-Forest-Modell, mit dem die besten Ergebnisse erzielt wurden, wurde schließlich zur automatisierten Extraktion impactrelevanter Textpassagen eingesetzt. Die Outputs des Modells wurden mithilfe verschiedener Heuristiken (bspw. des Entfernens von Bildunterschriften, Inhaltsverzeichnissen, Aufzählungen etc.) semi-automatisiert bereinigt.

3.1.3.4.2. Manuelle Annotation der Projektberichte

Anschließend wurden die Daten für die Einspeisung in das Annotationstool INCEpTION[32] aufbereitet. INCEpTION, dessen Entwicklung von der DFG gefördert wurde und der Creative-Commons-Lizenz unterliegt, unterstützt die serverbasierte kollaborative manuelle Annotation von Texten auf beliebig vielen Ebenen (Layers) mit flexibel und inkrementell erweiterbaren Annotationsschemata (Tagsets) und war daher für die Annotation von Impactkategorien auf der Grundlage eines eigens entwickelten Impact-Kategorien-Schemas ideal geeignet.

[30] https://github.com/google-research/bert
[31] https://github.com/facebookresearch/fairseq/tree/main/examples/roberta
[32] https://inception-project.github.io/

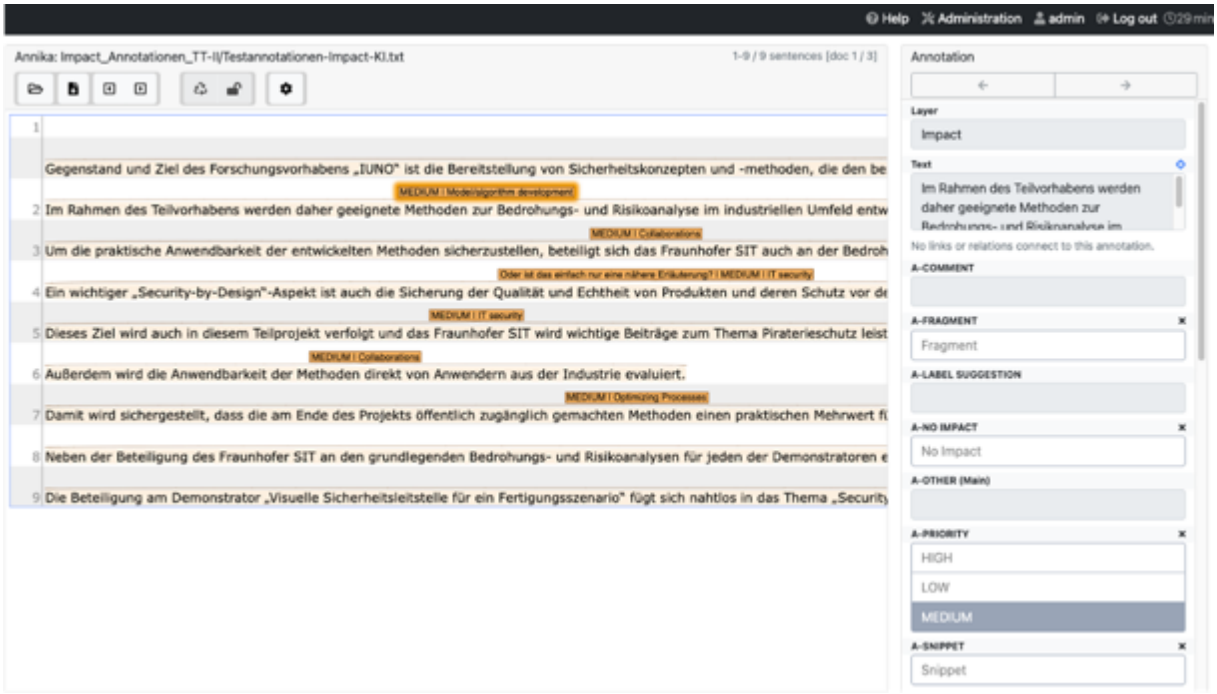

Abbildung 3-4 Screenshot INCEpTION mit dem TextTransfer Custom-Tagset.

Hierfür wurde zunächst ein Custom-Tagset mit allen im Codebook definierten Annotationskategorien angelegt. Anschließend wurden die Annotierenden im Rahmen von drei Trainingssitzungen in ihre Aufgaben eingearbeitet. Die Trainingssitzungen beinhalteten dabei eine Vorstellung des Projekts *TextTransfer II* und des Themas Impact in wissenschaftlichen Texten, eine Einführung in das Codebook, eine Einführung in das Annotationstool INCEpTION sowie Probeannotationen und deren Besprechung.

Die Annotationen wurden anschließend auf Satzebene in einem zweistufigen Prozess durchgeführt:

Die erste Phase diente vor allem der Validierung und Feinabstimmung des Annotationsmanuals, der Verbesserung der Übereinstimmung der Annotierenden und der Verbesserung der Qualität und Zuverlässigkeit der Annotationen. In dieser Phase wurde eine ausgewählte Teilmenge von 3.000 Sätzen aus allen vier Domänen unabhängig voneinander von zwei Annotierenden pro Satz annotiert. Die Daten wurden hierbei in drei Pakete mit jeweils 1.000 Sätzen aufgeteilt, nach Abschluss jedes Pakets wurden die Annotationen in Datensitzungen besprochen. Bei Bedarf wurden die Richtlinien optimiert bzw. verfeinert, indem konkretere Anweisungen oder Beispiele hinzugefügt wurden. Nach Abschluss

jedes Pakets wurde die Übereinstimmung der Annotationen mithilfe der Inter-Annotator-Agreement-Metrik Cohen's Kappa gemessen, welches sowohl die beobachtete als auch die zufällige Übereinstimmung der Annotierenden berücksichtigt. Dabei konnte eine kontinuierlich wachsende Übereinstimmung festgestellt werden:

Cohen-Kappa-Score Textpaket 1	Cohen-Kappa-Score Textpaket 2	Cohen-Kappa-Score Textpaket 3
34,68	52,83	74,48

Abbildung 3-5 Cohen-Kappa-Score Textpakete 1 bis 3.

Der Cohen-Kappa-Score für das erste Textpaket betrug 34,68 und stieg für das zweite Paket auf 52,83. Beim letzten Paket konnte schließlich ein hohes Maß an Übereinstimmung mit einem Kappa-Score von 74,48 festgestellt werden. Dies ließ sich als Indikator dafür werten, dass das Annotationsschema geeignet war, um den Impact in Projektberichten aus verschiedenen Domänen angemessen und eindeutig zu erfassen und zu kategorisieren. Es zeigte auch, dass die Anweisungen in dem Codebuch ausreichend beschrieben und dargestellt waren. Sätze mit voneinander abweichenden Annotationen aus dieser ersten Annotationsphase wurden schließlich von einem Expert-Annotator beurteilt und nachannotiert.

Nachdem ein hohes Maß an Übereinstimmung erreicht wurde, wurden anschließend in der zweiten Annotationsphase die verbleibenden Dokumente von jeweils einem trainierten Annotierenden annotiert. Die Kombination dieser beiden Annotationsansätze ermöglichte es, die Qualität und Zuverlässigkeit der Annotationen sicherzustellen und bot eine effiziente Möglichkeit, große Mengen von Texten manuell mit Impactkategorien zu kodieren.

Die annotierten Berichte wurden sukzessive auch an den Unterauftragnehmer UIUC für die Arbeiten aus AP 4 Maschinelles Lernen übergeben.

3.1.3.5. Identifikation externer Impact-Referenzen

Wie bereits im Vorangegangen erwähnt war die Identifikation neuer externer Impact-Referenzen zur Umsetzung des Top-down-Ansatzes aus der ersten Förderphase des Projekts ein Anliegen von

TextTransfer II. Um die hohen Aufwände, die bei der Etablierung stabiler Aussagen über die Impact generierende tatsächliche Verwertung von Forschungsergebnissen mittels Interviews anfallen, für die Methodennutzer zu minimieren, waren in *TextTransfer II* Wege zur effizienteren Erschließung von Metadaten durch den Projektpartner IDS zu evaluieren. Datengrundlage hierfür bot das am IDS entwickelte Deutsche Referenzkorpus (DeReKo)[33], das mit über 57,6 Milliarden Wörtern (zum Zeitpunkt der Berichtserstellung) das weltweit größte Korpus mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit darstellt. Um zu eruieren, inwiefern es der Wissenschaft gelingt, ein Thema im öffentlichen Diskurs zu platzieren, wurde nach medialen Texten gesucht, die über die Projekte aus dem Datensatz (Quelltexte) berichten. Die zugrundeliegende Idee hierbei war, dass mediale Texte (also bspw. Zeitungsberichte) als geeignete Indikatoren für das Aufspüren bzw. Messen von Impact auf öffentliche Diskurse fungieren können.

In einer ersten Pilotstudie zu den Projektberichten der geisteswissenschaftlichen Domänen Germanistische Linguistik und Musikwissenschaften wurden hierfür verschiedene Kombinationen von Suchausdrücken angewendet. Mittels mehrteiliger Suchausdrücke konnten im Rahmen dieser Pilotstudie für diese beiden Disziplinen keine relevanten Treffer erzielt werden, allerdings erwies sich die Suche nach ausschließlich dem Projektnamen oder den Projektbeteiligten als zielführender. Bei der Durchsicht der erzielten Treffer zeigten sich jedoch einige Ambiguitäten: Da viele der Projektnamen Akronyme sind, die gleichzeitig bedeutungstragende Wörter darstellen (bspw. Seelen, Kobra, emergent, Salut), führte die Suche hier zu vielen False Positives, also fälschlicherweise detektierten Zeitungsberichten, die also nicht über die gesuchten Projekte berichteten. Als eine weitere Ursache für False Positives stellte sich der Umstand heraus, dass die Namen der Projektbeteiligten teilweise häufig vorkommende Namen im Deutschen sind (bspw. Bernd Müller).

Für die weitgehend technikorientierten Domänen Mobilität und Künstliche Intelligenz wurden in einer zweiten Pilotstudie im weiteren Projektverlauf einfache Suchanfragen auf der Grundlage der Projektnamen durchgeführt, mittels derer mehr Treffer erzielt werden konnten als bei den geisteswissenschaftlichen Domänen. Dieses vorläufige Ergebnis wurde durch eine systematische Recherche und

[33] https://www.ids-mannheim.de/digspra/kl/projekte/korpora/

komplexe Suchausdrücke mittels der KorAP API (Application Programming Interface)[34], die anhand von Python-Scripts auf das DeReKo zugreift und so eine umfangreichere und effizientere Suche ermöglicht, überprüft.

Hierfür wurden Tabellen mit relevanten Metadaten für alle vier Domänen erstellt, die als Grundlage für die Bildung komplexer Suchausdrücke fungieren. Diese Tabellen beinhalteten neben den Projektnamen die projektbeteiligten Personen und Institutionen, die Projektlaufzeit sowie eine Kurzbeschreibung des Projekts.

Abbildung 3-6 Relevante Metadaten: Screenshot aus einer Tabelle der Domäne Linguistik.

Aus der Kurzbeschreibung des Projekts wurden mittels eines Python-Scripts alle Inhaltswörter extrahiert, die ebenfalls zur Bildung der komplexen Suchausdrücke herangezogen wurden.[35] Des Weiteren konnte durch die Pilotstudien eruiert werden, dass sich bei den Namen der Projektbeteiligten nur die Nachnamen als zielführende Komponenten der Suchausdrücke eignen.

Da die Bildung komplexer Suchanfragen bestehend aus Metadateninformationen über die Projekte wie bspw. der Projektname, die beteiligten Personen oder Institutionen zwar zu einer hohen Präzision, jedoch einer niedrigen Abdeckung (Recall) der Suchanfragen führte, wurde letztendlich ein Workflow bestehend aus der Suche nach Projektnamen und anschließender manueller Filterung der Ergebnisse

[34] https://korap.ids-mannheim.de/doc/api

[35] Das Skript ist am IDS gespeichert und kann bei Bedarf per Email an bopp@ids-mannheim.de angefordert werden.

entwickelt. Auf diese Weise konnten insgesamt 412 Zeitungstexte (*True Positives*) gefunden werden, in denen über die Projekte aus dem *TextTranfer*-Datensatz berichtet wird. Für 22% der Projekte wurde mindestens ein Zeitungstext gefunden, wobei über viele Projekte auch mehrfach berichtet wurde. Die durchschnittliche Anzahl an Zeitungsberichten pro Projekt lag bei 5,5.

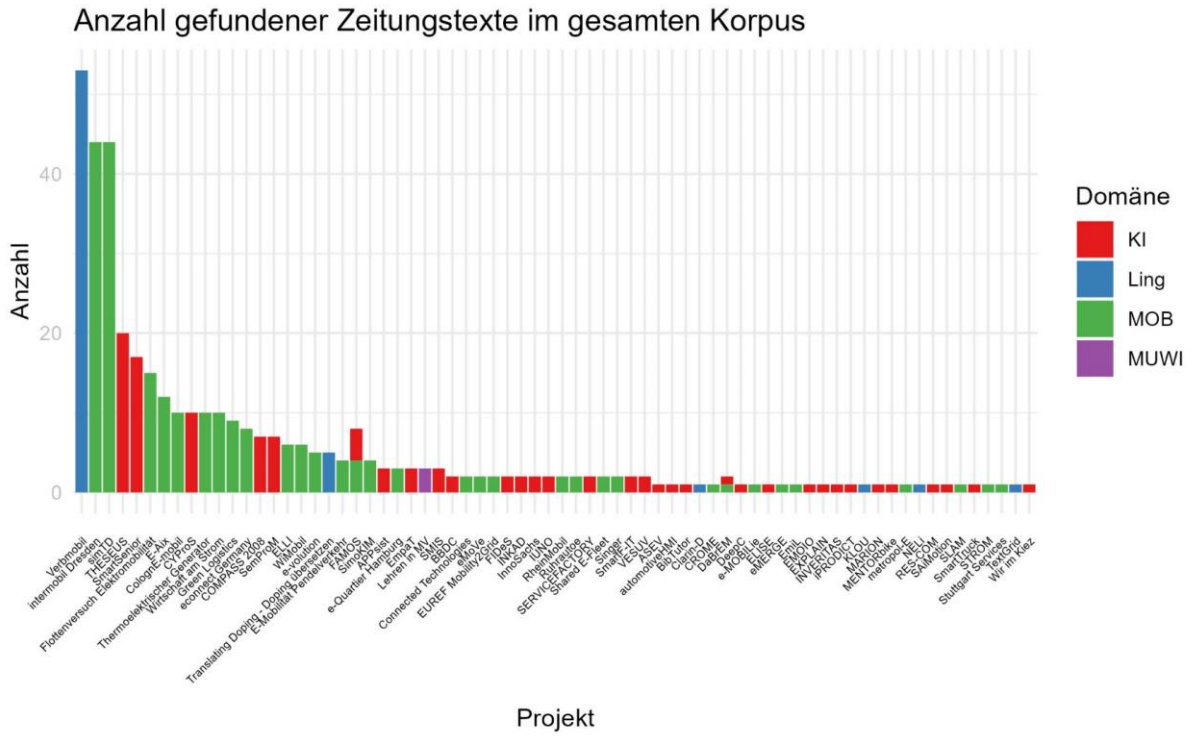

Abbildung 3-7 Extrahierte Berichterstattungen pro Projekt aus dem Deutschen Referenzkorpus.

Basierend darauf wurde eine Datenbank erstellt, in der die Projektberichte und die dazugehörigen Zeitungstexte verknüpft wurden. Des Weiteren wurden die Texte mit vielfältigen automatischen und manuellen Annotationen und Metadaten angereichert, um umfassende korpuslinguistische Analysen zu ermöglichen.

Folgende Angaben wurden automatisiert erfasst bzw. aus den Metadaten des Deutschen Referenzkorpus extrahiert:

FKZ 01IO2002A
Version 1.1

- Schlüsselwörter → automatisierte Extraktion aus den Texten mittels eines Python-Skripts
- Frequente Mehrwortverbindungen → automatisierte Extraktion aus den Texten mittels eines Python-Skripts
- Eigennamen → automatisierte Extraktion aus den Texten mittels eines Python-Skripts
- Textlänge (Anzahl der Tokens) → automatisierte Ermittlung mittels eines Python-Skripts
- Land der Zeitung (Deutschland, Österreich, Schweiz) → Extraktion aus den Metadaten des Deutschen Referenzkorpus)
- Art der Zeitung (Tageszeitung, Wochenzeitung, Fachzeitschrift, Publikumszeitschrift, Wochenzeitschrift, Monatszeitschrift) → Extraktion aus den Metadaten des Deutschen Referenzkorpus
- Textdomäne (z. B. Panorama, Politik, Lokales) → Extraktion aus den Metadaten des Deutschen Referenzkorpus
- Publikationsort → Extraktion aus den Metadaten des Deutschen Referenzkorpus
- Publikationsdatum → Extraktion aus den Metadaten des Deutschen Referenzkorpus

Folgende Angaben und Kategorien wurden manuell (mittels Annotationen durch Hilfskräfte und Projektmitarbeitende) identifiziert:

- Anteil der Passagen mit Projektbezug aus den Zeitungstexten (Prozentangabe mit den Optionen 0, 5, 25, 50, 75, 100)
- Textsorte (Nachrichten, Artikel (Projekt), Artikel (Person), Interview, Kommentar, Studien, Erklärartikel, Veranstaltungsberichte, Buchrezensionen, Wissenschaftsgeschichten)
- Zeitungstyp (regional, überregional, Fachzeitschrift)

Die Angaben wurden anschließend in eine Datenbank überführt, um mögliche Korrelationen zwischen den Parametern zu eruieren und auf diese Weise festzustellen, welche Faktoren es begünstigen, dass über ein Projekt in Online- und Printmedien berichtet wird.

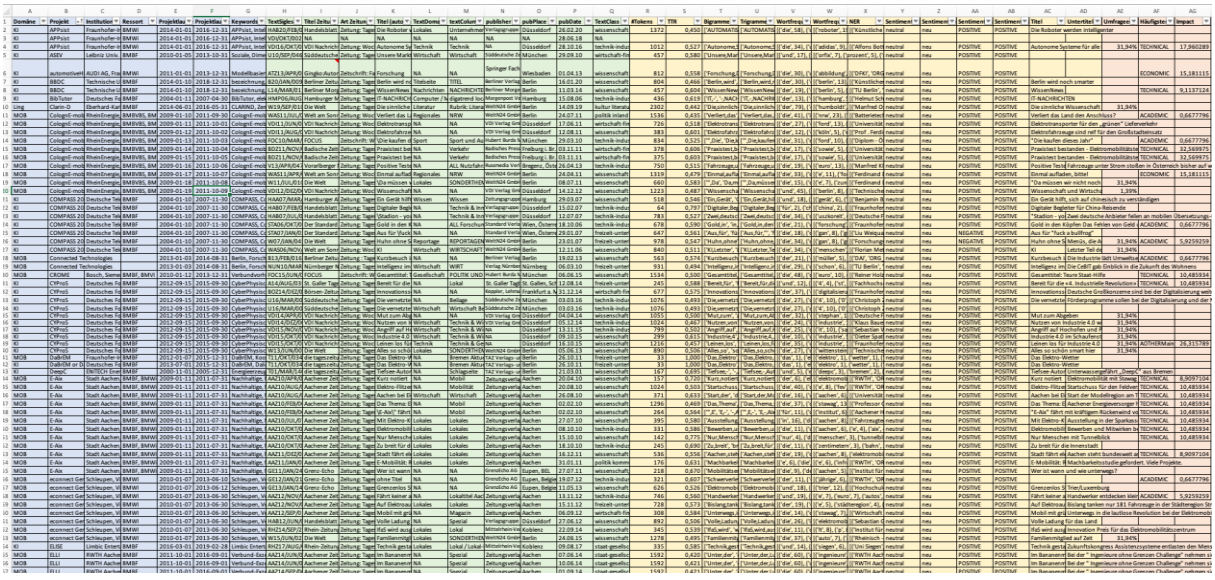

Abbildung 3-8 Erfassung (linguistischer) Merkmale von Zeitungsberichten über wissenschaftliche Forschungsprojekte.

Dabei konnte u.a. gezeigt werden, dass insbesondere in den Zeitungsrubriken Wissenschaft, Reportage und Computer über die Projekte des genannten Datensatzes berichtet wird. Ein Großteil der Berichte widmen sich dabei entweder gänzlich den Forschungsprojekten oder erwähnen diese nur am Rande. Die Berichterstattungen erfolgen dabei in den meisten Fällen ab der zweiten Hälfte der Projektlaufzeit, wobei über einige Projekte auch weit nach Projektabschluss noch berichtet wird (z. B. über das Projekt Verbmobil oder simTD).

FKZ 01IO2002A
Version 1.1

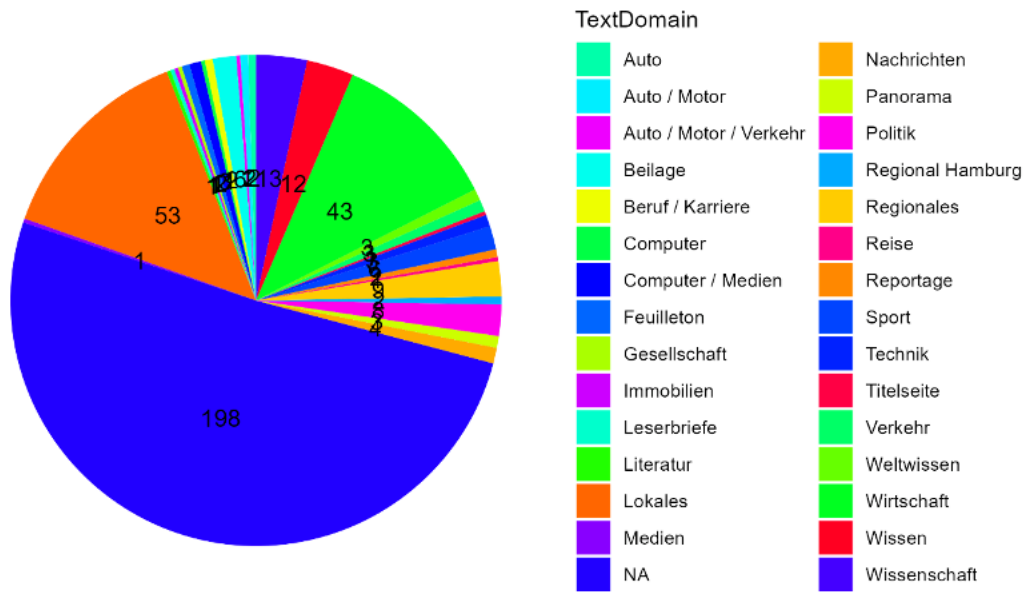

Abbildung 3-9 Zeitungsrubriken aller gefundenen Projekte im Deutschen Referenzkorpus.

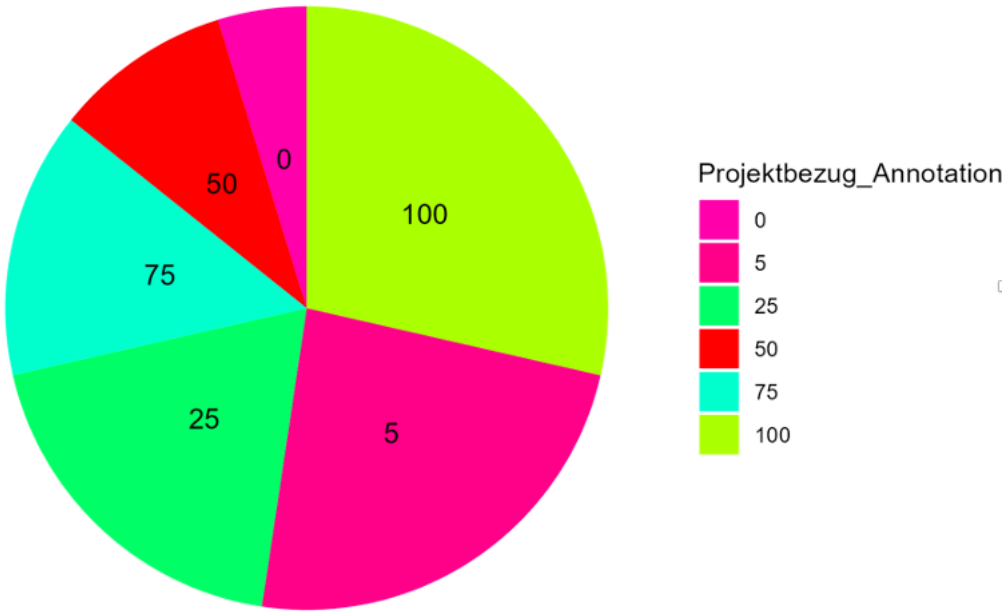

Abbildung 3-10 Projektbezug aller gefundenen Projekte im Deutschen Referenzkorpus.

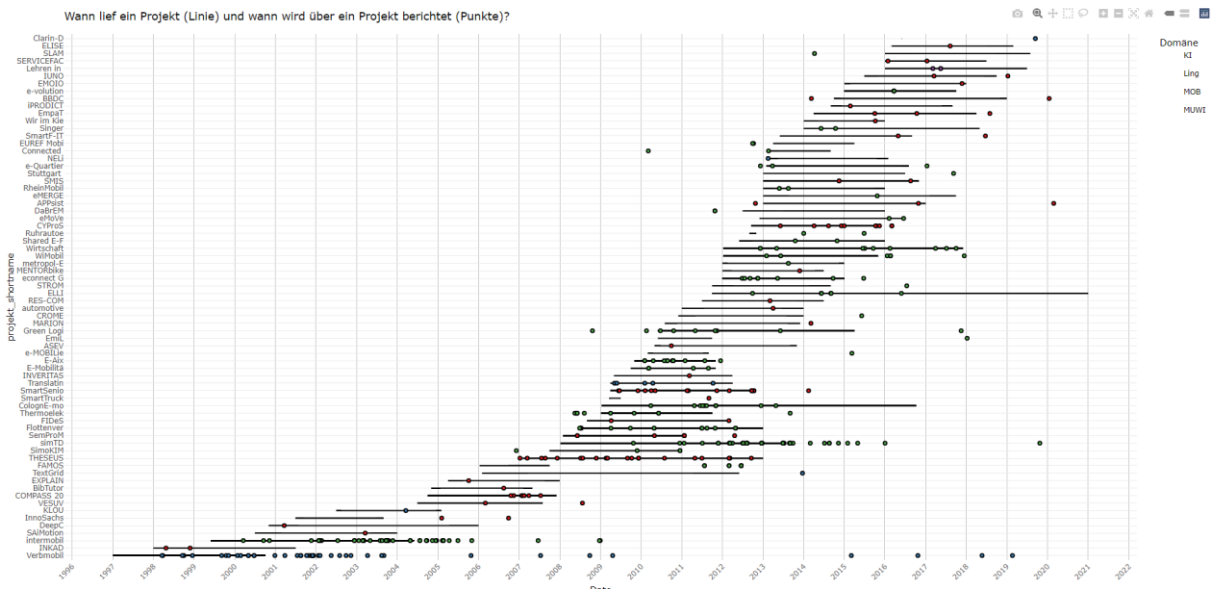

Abbildung 3-11 Zeitpunkte der Berichterstattung.

Erfasst wurden in der Datenbank auch die Impact-Annotationen, um auf diese Weise zu eruieren, inwiefern bestimmte Impactkategorien die Chance auf eine Berichterstattung über ein Projekt begünstigen, und somit gleichzeitig die Teilprojekte des *TextTransfer*-Projekts zusammenzuführen. Dabei konnte gezeigt werden, dass insbesondere über Projekte medial berichtet wird, in denen laut Projektbericht der technische Impact im Vordergrund steht, während Projektberichte, die vorwiegend den akademischen Impact eines Forschungsprojekts betonen, tendenziell unterrepräsentiert in der medialen Berichterstattung sind.

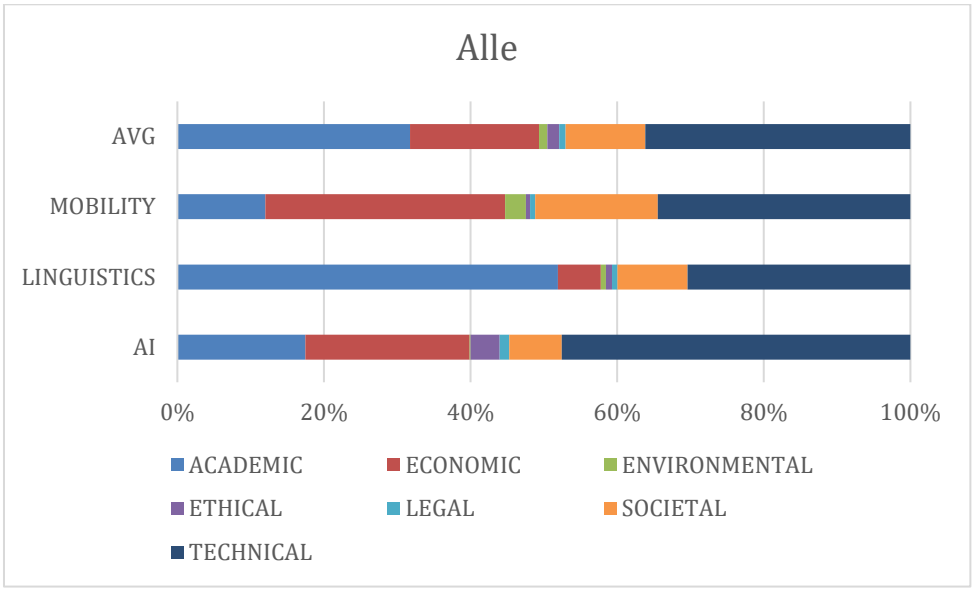

Abbildung 3-12 Verteilung der Impactkategorien im Gesamtdatensatz.

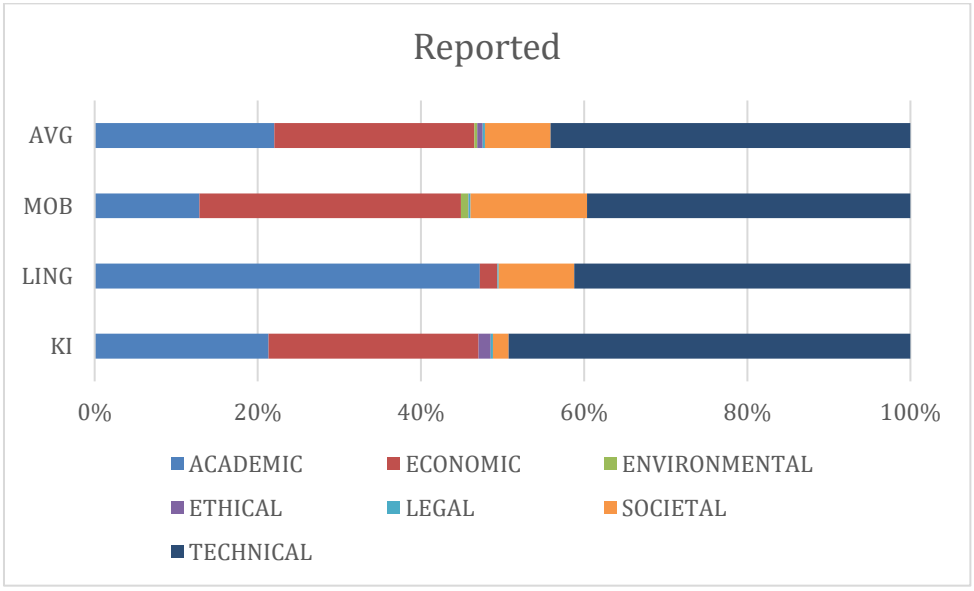

Abbildung 3-13 Verteilung der Impactkategorien in den Projektberichten, über die medial berichtet wurde.

FKZ 01IO2002A
Version 1.1

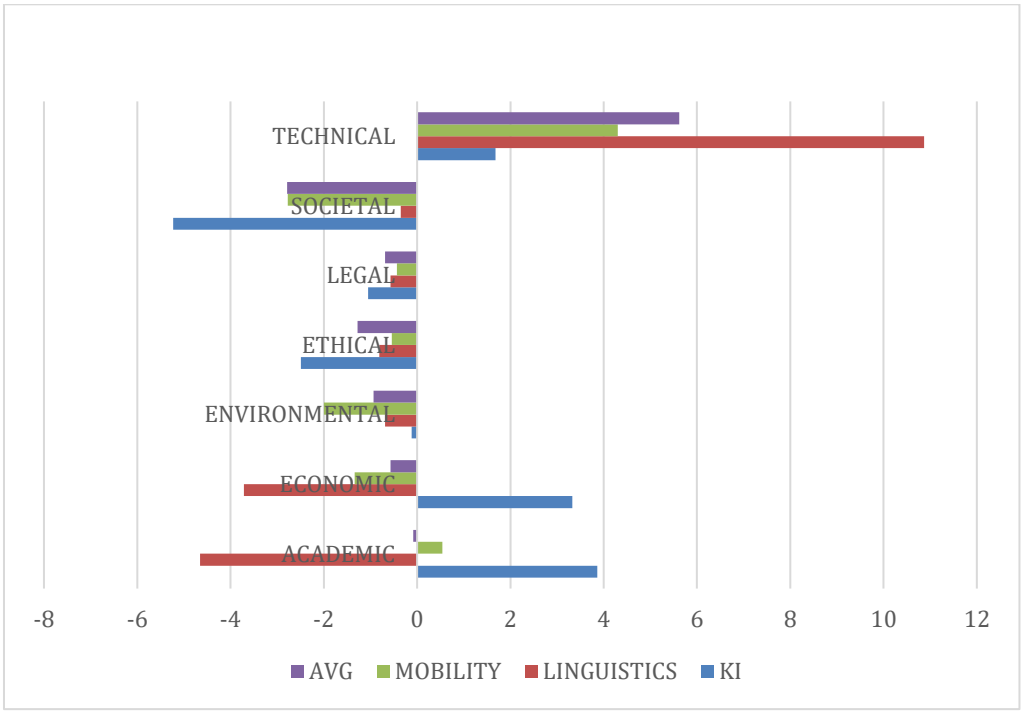

Abbildung 3-14 Differenz zwischen den Impactkategorien in den Projektberichten, über die medial berichtet wurde, und dem Gesamtdatensatz in Prozentpunkten.

AP3 war zu Projektende planmäßig abgeschlossen.

3.1.4. AP4: Maschinelles Lernen

Die in *TextTransfer I (Pilot)* mittels maschinellen Lernens auf die Analyse von Projektendberichten der Domäne „Mobilität“ spezialisierte Methode sollte im Hauptprojekt durch den Projektpartner IDS – mit Unterstützung des Unterauftragnehmers UIUC – auf die Anforderungen einer thematisch wie formal heterogenen Datenbasis trainiert und angepasst werden – mit dem Ziel der Entwicklung maschineller Lernverfahren, die in Texten automatisch Sätze bzw. Abschnitte detektieren, die das Impactpotenzial eines Textes ausdrücken und nach diesem Impact klassifizieren.

Mithilfe der neu hinzugekommenen Domänen Künstliche Intelligenz, Germanistische Linguistik und Musikwissenschaften neben der bereits vorhandenen Domäne Mobilität wurde hierfür an der Vertiefung der Spezifikation des Maschinellen Lernens gearbeitet. Projektendberichte aller vier Domänen waren wie bereits erwähnt im Vorfeld vom Projektpartner TIB extrahiert und konvertiert worden und anschließend durch das IDS als Stichproben für die Verwendung beim Maschinellen Lernen auf- und vorbereitet worden.

Wie ebenfalls in diesem Kapitel unter AP 3 bereits dargelegt wurde hierfür zunächst auf der Grundlage des bottom-up-Annotationsprozesses ohne vordefinierte Kategorien ein Annotationsschema mit verschiedenen Haupt- und Unterkategorien von Impact entwickelt, die sich auf die Projektberichte aller vier Domänen anwenden ließen.

Bis zum Ende der Projektlaufzeit wurden insgesamt 65.000 Instanzen (Sätze) mit Ober- und Unterkategorien von Impact annotiert. Der annotierte Datensatz wurde anschließend zunächst statistisch ausgewertet, bspw. hinsichtlich des Anteils impactrelevanter Sätze, der ebenfalls annotierten Kategorie der Impactintensität (Hoch - Mittel - Niedrig), der Verteilung der einzelnen Impactkategorien pro Domäne, sowie ko-okkurierender Impactkategorien. Dabei zeigte sich u.a., dass die Verteilung der Kategorien zwischen den einzelnen Domänen teils deutlich divergierten: So wiesen Berichte aus den Domänen Künstliche Intelligenz oder Mobilität erwartungsgemäß mehr Sätze auf, die auf technischen Impact verweisen, während in den Domänen Musikwissenschaften oder Linguistik mehr Sätze enthalten waren, die gesellschaftlichen bzw. kulturellen Impact ausdrückten. Andere Kategorien wie etwa ethischer Impact kam nur bei einigen Projekten, jedoch domänenübergreifend vor, während er bei anderen Projektberichten gar nicht zu finden war. Rechtlicher bzw. politischer Impact war in der Pilotstudie die am seltensten vergebene Kategorie.[36]

Der Datensatz, der, wie bereits beschrieben, mit verschiedenen Ober- und Unterkategorien von Impact annotiert wurde, wurde schließlich zum Training maschineller Lernverfahren verwendet. Hierbei wur-

[36] Weitere Details zur statistischen Auswertung des annotierten Datensatzes finden sich in der Anlage.

den (supervised) Lernverfahren entwickelt, trainiert und evaluiert, um auf diese Weise die automatische Vorhersage von Impact wissenschaftlicher Forschung anhand von Projektberichten zu ermöglichen. Bei supervisierten Lernverfahren werden annotierte Daten verwendet, um Algorithmen zu trainieren, die Daten klassifizieren oder Ergebnisse genau vorhersagen. Wenn Eingabedaten in das Modell eingespeist werden, werden die Modellparameter so lange automatisiert angepasst, bis korrekte Zielvorhersagen gemacht werden. Da während der Projektlaufzeit verschiedene Large Language Models (LLM/Sprachmodelle) veröffentlicht bzw. verfügbar gemacht wurden, wurden diese neuen Technologien in den Experimenten ebenfalls eingesetzt.

Während „herkömmliche“ neuronale Netze (z. B. rekurrente neuronale Netzwerke) Eingaben sequentiell verarbeiten, verarbeiten Large Language Modelle sie simultan in Blöcken, in denen mittels des sogenannten “Attention”-Mechanismus die semantischen und syntaktischen Merkmale erfasst und automatisiert wichtige Wörter bzw. Sequenzen in Sätzen und Texten erkannt werden können. Transformer Language Modelle werden auf sehr großen Datenmengen vortrainiert (bspw. wurde das RoBERTa-Modell auf 160 GB an Nachrichten, Büchern, Geschichten und Webtexten vortrainiert) und können anschließend für verschiedene andere Aufgaben adaptiert werden, wie beispielsweise die Klassifikation von Sätzen – im vorliegenden Fall die Zuweisung einer bestimmten Impact-Kategorie für einen Satz. Im Rahmen von *TextTransfer II* wurden die Transformer Modelle BERT, Llama und GPT[37] angewendet. Alle drei Modelle sind zur Zeit der Berichterstellung State-of-the-Art Modelle zur automatischen Sprachverarbeitung in der Computerlinguistik und sind, wie in zahlreichen Studien gezeigt werden konnte, in besonderer Weise für die Klassifikation von Sätzen geeignet.

Die genannten Modelle wurden eingesetzt, um den potenziellen Impact von Forschungsprojekten automatisch zu prognostizieren und zu klassifizieren. Dabei wurden sowohl Finetuning-Ansätze als auch Prompting bzw. Zero- oder Few-Shot Learning erprobt. Hierbei handelt es sich um verschiedene Methoden, um die Modelle für bestimmte Aufgaben (hier: die Prognose von Impactkategorien) zu nutzen und anzupassen. Beim Finetuning wird ein vortrainiertes Modell auf einem spezifischen Datensatz (im vorliegenden Fall die mit Impactkategorien annotierten Projektberichte) weiter trainiert, um es für

[37] https://huggingface.co/docs/transformers/index

FKZ 01IO2002A
Version 1.1

eine bestimmte Aufgabe zu optimieren. Dieser Prozess ermöglicht es dem Modell, sich an die Besonderheiten und Feinheiten der neuen Daten anzupassen. Zero-Shot-Verfahren (auch prompt engineering) beziehen sich auf die Fähigkeit eines Modells, Aufgaben zu bewältigen, für die es nicht explizit trainiert wurde. Bei diesen Verfahren erhält das Modell eine (möglichst präzise formulierte) Aufgabe, die es auf der Basis des im Modell bereits enthaltenen Wissens löst, ohne hierfür Trainingsdaten zu verarbeiten. In vorliegenden Fall wurde dem Modell eine Liste der Impactkategorien gegeben, kurze Erklärungen hierzu, und anschließend die zu klassifizierenden Texte. Beim Few-Shot-Lernen erhält das Modell zusätzlich zu den Instruktionen eine sehr kleine Anzahl von Beispielen (sogenannte „Shots“), um die gestellte Aufgabe zu lösen, im vorliegenden Fall einige annotierte Sätze aus dem Datensatz.

Die Ergebnisse zeigten, dass die Finetuning-Ansätze signifikant besser als Zero- oder Few-Shot Verfahren funktionierten, wohingegen neue Klassifikationsschemata eine Herausforderung für vortrainierte (aber nicht finegetunte) LLMs wie ChatGPT sind: So konnten bspw. mit dem finegetunten BERT-Modell Accuracy-Werte von bis zu 72% bei der Vorhersage der Hauptkategorien erzielt werden, während ChatGPT beim Zero-Shot-Learning nur 53% und beim Few-Shot-Learning nur 54% Akkuratheit erzielte. Dies verdeutlicht die hohe Relevanz von annotierten Trainingsdaten für den Task der automatischen Impactprognose.

Darüber hinaus wurde in einer Versuchsreihe der Zusammenhang des Impacts der Forschungsprojekte mit Sentimentklassen erforscht. Bei solchen Sentiment-Analysen („Stimmungserkennung“) erfolgt eine automatische Auswertung von Texten mit dem Ziel, textuelle Einheiten als positiv, negativ oder neutral zu klassifizieren. Die Grundannahme ist dabei, dass in wissenschaftlichen Textprodukten wie Abschlussberichten eigentlich neutral geschrieben werden sollte. Wenn dem nicht so ist, wenn also ein positives oder negatives Sentiment vorliegt, dann wird davon ausgegangen, dass dies zu einem bestimmten Zweck geschieht – zum Beispiel, dass (positiver) Impact ausgedrückt werden soll. Daher wurden mögliche Korrelationen von Impactklassen und Impactintensitäten – beides manuell annotiert – mit den Sentimentklassen positiv und negativ untersucht. Es wurden verschiedene automatisierte

Sentimentmodelle (SentiWS, SentiMerge, SentiBERT, GerVADER, Fast Text und MONAPipe) für die Projektberichte adaptiert, angewendet und evaluiert, wobei sich SentiBERT und GerVADER als akkurateste Modelle für die im Projekt verwendeten Daten zeigten.

Die quantitativen und qualitativen Analysen der Ergebnisse zeigten nicht nur, dass die Projektberichte nicht völlig neutral sind, sondern auch positive Anteile aufweisen. Darüber hinaus konnte durch linguistische und statistische Analysen eine Korrelation zwischen positiver Polarität und Impact, insbesondere bei Sätzen mit hoher Impactintensität, gezeigt werden. Dies deutet darauf hin, dass positiv bewertete Wörter in Projektberichten verwendet werden, um die Ergebnisse der Forschungsprojekte hervorzuheben.

AP4 war zu Projektende planmäßig abgeschlossen.

3.1.5. AP5: Technische und rechtliche Rahmenbedingungen

Aus den Vorarbeiten von *TextTransfer I (Pilot)* ergaben sich Bedarfe insbesondere hinsichtlich technischer und rechtlicher Rahmenbedingungen, um die Anwendung maschineller Lernverfahren für textuelle Stichproben, aber auch im Speziellen bzgl. der Nutzung der Methode *TextTransfer* zu optimieren bzw. zu verstetigen. Zum Zeitpunkt der Projektdurchführung existiert kein technischer oder formaler Standard bei der Einreichung öffentlich geförderter Projektabschlussberichte. Dies betrifft vornehmlich das digitale Format, die Vergabe von geeigneten Metainformationen sowie rechtssichere Bestimmungen zur Nutzung von Abschlussberichten. Diese Hemmnisse stehen einer künftigen maschinellen Auswertung und Nutzung im Wege. Infolgedessen können viele von der öffentlichen Hand geförderten Drittmittelvorhaben keiner dem Gehalt der Ergebnisse angemessenen, mit zeitgemäßen Methoden erschlossenen Nutzung zugeführt werden.

3.1.5.1. Standardisiertes Abgabeformat

Im Pilotprojekt wurde der Bedarf nach einer optimierten Vorbereitung insbesondere des Quelltyps Projektbericht durch ein standardisiertes Abgabeformat identifiziert, um die erheblichen Aufwände einer für Maschinenlesbarkeit notwendigen Datenkonvertierung künftig zu umgehen. An der TIB

wurde entsprechend antragsgemäß ein prototypisches Tool entwickelt, mithilfe dessen Projektendberichte – der IDS-Abschlussbericht von *TextTransfer I* fungierte dabei als Beispiel für den Quelltyp Projektbericht– nach offenen, ausgereiften Dokumentenstandards (wie z. B. Markdown und XHTML) erstellt und in einem maschinenlesbaren Format eingereicht werden können. *TextTransfer II* hatte sich hierbei, soweit möglich und vorhanden, nach den Vorlagen der Fördermittelgeber gerichtet. Die TIB hatte hierzu die bereits genutzte Publishing Pipeline Fidus Writer auf der Plattform Github verwendet und angepasst[38], um einen Texteditor für strukturierten Text mit Funktionen zur Erstellung von Templates zur kollaborativen Bearbeitung von Berichten sowie zum Speichern von Überarbeitungen und Metadaten zu entwickeln. Bezugnehmend auf Anforderungen aus *TextTransfer II* hatten die Entwickler von Fidus Writer bereits fehlende Features in der Software nachgerüstet.[39] Künftig wird auch eine automatisierte Einreichung in das DSpace-Repository der TIB vorgesehen sein. Vorbereitend hierfür hatte die TIB bereits ein neues Dspace-Repository unter https://oa.tib.eu/renate/ aufgesetzt; ein entsprechender Submission-Workflow wurde hierfür entwickelt. Für single-sign-on-Verfahren wurden mögliche Lösungen wie Github (kommerziell), DFN AAI oder Overleaf evaluiert.

Ferner wurde an der TIB ProposalPilot entwickelt.[40] ProposalPilot ist ein Demonstrator (Technologie-Reifegrad 6), der zeigt, wie mit einem besonders schlanken (die Umsetzung erfolgt ausschließlich mit PHP sowie JavaScript, d.h. ohne eigene Datenbank) technischen Ansatz

- strukturierte Forschungsberichte in einem benutzerfreundlichen, barriere-armen Webformular erfasst,
- in JSON, HTML, PDF sowie Metadaten in XML nach den Dublin-Core-Spezifikationen von DSpace 7 ausgegeben,
- sowie optional zur realitätsnahen Nachnutzbarkeit (den Prinzipien von FAIR Data folgend) als Issue in einem frei gewählten GitHub-Repository als Issue[41] gespeichert werden kann.

[38] https://github.com/TIBHannover/Fidus-Writer-Manual
[39] https://www.fiduswriter.org/2021/07/29/fidus-writer-3-10-with-folders-and-downloadable-document-templates/.
[40] Vgl. den Open-Source-Quellcode des Projekts unter https://github.com/TIBHannover/text-transfer-ii-prototype, sowie den Demonstrator live unter https://proposalpilot.texttransfer.org/
[41] https://docs.github.com/de/issues

Fördergeber oder andere Organisationen, die von Proposalpilot ausgehend einen Prozess zur Einreichung strukturierter Forschungsberichte einrichten, könnten hier sehr einfach

- eigene Metadatenfelder in ihrem Eingabeformular ergänzen,
- eine Authentifizierung durch Token ergänzen,
- dem ProposalPilot lokale Schreibrechte auf dem eigenen Server geben. (Letztere Funktion ist im Demonstrator aus Sicherheitsgründen deaktiviert.)

3.1.5.2. Rechtliche Rahmenbedingen

Basierend auf den Erkenntnissen aus dem Pilotprojekt war in AP 5 außerdem der Frage nachzugehen, inwieweit die rechtlichen Rahmenbedingungen eine Verstetigung der Nutzung der Methode unterstützen. Die Expertise, die zur Beantwortung dieser Frage notwendig ist, ist mit den Kapazitäten des Wissens zum juristisch und ethisch stabilen Umgang mit digitalen Forschungsdaten bereitstellenden CLARIN Legal Helpdesk als besondere Expertise des IDS vorhanden.[42] Rechtliche Aspekte greifen immer dann, wenn ein Verfahren des Data Mining – *TextTransfe*r zählt unter Verwendung maschineller Lernverfahren zu diesem Bereich – auf eine Datengrundlage – seien es Quelltypen, seien es Referenzdaten – zugreift, die von Rechten Dritter belegt sind. Als wesentliche rechtliche Hürden sind hierbei einerseits Lizenzierungen von Daten zu nennen, andererseits die Verarbeitung personenbezogener Daten.

Bezüglich der Quelltypen hat sich *TextTransfer* im Wesentlichen auf öffentliches Material gestützt, das als Belegexemplar beim Projektpartner TIB hinterlegt wurde. Die Erhebung anderer Formate – zu denken wäre insbesondere an Antragstexte – scheiterte hingegen an strategischen und rechtlichen Ursachen. Im Sinne der Generalisierung der Methodennutzung strebt *TextTransfer* danach, einen möglichst barrierefreien Zugang zu Quelldaten vorauszusetzen. Personenbezogene Daten finden sich dabei schwerpunktmäßig in den zur Klassifizierung von Quelltexten herangezogenen Referenzdaten. Eine Verarbeitung und (wenn auch nur projektinterne) Weitergabe dieser Informationen unterliegt somit grundsätzlich dem Datenschutz, dem das Teilprojekt durch eine Reihe von Maßnahmen gerecht wurde.

[42] https://www.clarin-d.net/de/konferenz-abstracts/405-der-clarin-d-helpdesk

Für dieses Szenario galt es somit, zwei wesentliche Aspekte abzusichern:

- Wann gilt deutsches Recht (oder die Datenschutzgrundverordnung [DSGVO], die in gewisser Weise Teil des deutschen Rechts ist)?

  und

- Was sind die Rollen der verschiedenen Akteure in dem von ihnen beschriebenen Arbeitsablauf?

Für alle Fälle von in Deutschland erhobenen Daten mit personenbezogenem Inhalt greift die DSGVO[43]:

a) durch Einzelpersonen und Organisationen mit Sitz in der EU, auch wenn die Verarbeitung selbst im Ausland stattfindet, z. B. auf asiatischen Servern (da das IDS in der EU ansässig ist, gilt die Datenschutz-Grundverordnung immer für die Aktivitäten dieser Instanz);

und

b) durch Einzelpersonen und Organisationen mit Sitz außerhalb der EU, sofern die Verarbeitung mit der Bereitstellung von Waren und Dienstleistungen für Menschen in der EU (z. B. einem Online-Shop) oder der Überwachung des Verhaltens von Menschen in der EU zusammenhängt.

Für die Verarbeitung solcher Daten hat das Projekt mithin sichergestellt, dass die Datenquelle über eine hinreichende eigene Berechtigung zur Bereitstellung dieser Daten verfügt oder die Erlaubnis der Urheber nachweisen kann.

Das IDS hat entsprechend alle personenbezogenen Daten anonymisiert; es wurde dabei sichergestellt, dass eine nachträgliche – durch vollständige Löschung (nicht bloße Trennung von den Daten) aller Indikatoren – Rekonstruktion technisch nicht mehr möglich war. Die DSGVO greift dann nicht mehr.[44]

[43] https://www.bmwi.de/Redaktion/DE/Artikel/Digitale-Welt/europaeische-datenschutzgrundverordnung.html
[44] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_de.pdf

Bei der Erhebung unter diesen Bedingungen ergab sich somit folgende, den technischen Workflow steuernde rechtliche Konstellation:

- Der Urheber ist ein unabhängiger Verantwortlicher für seine eigenen Zwecke und Mittel gem. Art. 4 Nr. 7 DSGVO;
- Die erhebende Instanz ist ein unabhängiger Verantwortlicher für seine eigenen Zwecke (es gibt keine gemeinsame Verantwortlichkeit IDS/Projektträger, da die beiden Stellen offenbar nicht gemeinsam darüber entscheiden, wie die Daten verarbeitet werden oder zu welchem gemeinsamen Zweck; jede von ihnen hat ihre eigenen Zwecke, für die sie autonom über die Mittel entscheidet);
- Verarbeitende Zwischeninstanzen (Auftragsverarbeiter) erhalten oder verarbeiten keine personenbezogenen Daten, so dass ihre Rolle im Sinne der DSGVO irrelevant ist; sie stellen lediglich beratend Fachwissen zur Verfügung, sind aber kein gemeinsamer Verantwortlicher für die anschließende Verarbeitung;

Besondere Beachtung fanden dabei die Status der verarbeitenden Zwischeninstanzen. Im vorliegenden Fall handelte es sich um die Unterauftragnehmer Görgen & Köller (G&K) und die Universität von Illinois, USA, (UIUC), die das Projekt bei der Datenanalyse unterstützten. Wie gezeigt wurde die rechtliche Absicherung der Urheber erreicht, indem ausschließlich nicht individuell zuordenbares Material übermittelt wurde. Ein Sonderfall trat ein, wenn die verarbeitende Instanz nicht-anonymisierte Daten vorhielt. Für die hier beschriebenen Zwecke trat dies in zwei Fällen auf:

a) Das IDS selbst erhob nicht-anonymisierte Daten. In diesem Falle war stets die Einwilligung des Urhebers einzuholen, welche die Zwecke der Erhebung zu wissenschaftlichen Zwecken, die Verpflichtung zur Anonymisierung bei Weitergabe und Veröffentlichung, die befristete Speicherung für die Projektlaufzeit bzw. für einen Zeitraum von 10 Jahren sowie die nachträgliche Revision der Einwilligung als wesentliche Eckpfeiler enthält. Intern setzte das IDS eine datenschutzkonforme Dokumentation (Verfahrensverzeichnis) auf, die den Erhebungsvorgang verbindlich festschrieb, den physischen Aufbewahrungsort umgrenzte sowie die Verpflichtung zur befristeten Aufbewahrung und

irreversiblen Anonymisierung festhielt. Der im IDS hauptamtlich tätige Datenschutzbeauftragte überwachte dabei kontinuierlich den gesamten Workflows und stellte die Einhaltung der Auflagen sicher.

b) Das vom IDS entwickelte Umfragewerkzeug *umfragewissen.texttransfer.org* hatte personenbezogene Daten auf freiwilliger Basis – insbesondere Kontaktadressen und Anstellungsverhältnisse – der Teilnehmenden abgefragt sowie obligatorisch IP-Adressen zur einmaligen Identifizierung von Probanden abgerufen. Auch hier wurden die Teilnehmenden im Vorfeld der Datenerhebung über Ziel und Zweck der Sammlung in Kenntnis gesetzt und eine Einwilligung zur Datenverarbeitung eingeholt. Die Bestimmungen zum Datenschutz wurden den Probanden dabei gesondert zur Kenntnis gebracht. Darüber hinaus wurde das Tool auf einem externen Serverdienst gehostet, mit dem ein gesonderter Vertrag zur Auftragsdatenverarbeitung abgeschlossen wurde, der den Vorgaben der DSGVO entspricht.[45]

Die Durchführung der o.g. Umfrage zur Tauglichkeit der im Projekt entwickelten Definitionen und Indikatoren zur Messung von Impact unter den Mitarbeiterinnen und Mitarbeitern des Projektpartners IDS wurde folgendermaßen angelegt, auf Basis des eigens mit dem Potenzial der Expansion zu einer vollwertigen Probandenplattform entwickelte Umfragetool, das im Rahmen dieser Umfrage einer ersten Probe unterzogen wurde. Auch hierfür charakterisierte die Art der Erhebungsdaten, vornehmlich deren Einstufung als personenbezogen, die rechtlichen Rahmenbedingungen der Durchführung.

Fielen bei einer Erhebung von Umfragedaten personenbezogene Angaben an, unterlagen Verarbeitung und (wenn auch nur projektinterne) Weitergabe grundsätzlich dem Datenschutz. [46]

Zur Durchführung der Erhebung wurde ein Verfahrensverzeichnis angelegt, um den Ablauf der Umfrage rechtstabil zu dokumentieren. Neben inhaltlichen Erhebungsdaten fielen weitere abgefragte Angaben an, die als personenbezogen eingestuft werden mussten. Zu den personenbezogenen Angaben zählten insbesondere Indikatoren, die dem Personalstamm der Probanden zuzurechnen waren:

[45] https://www.linode.com/de/legal-privacy/; https://www.linode.com/de/eu-model/
[46] https://www.bmwi.de/Redaktion/DE/Artikel/Digitale-Welt/europaeische-datenschutzgrundverordnung.html

- IP-Adresse
- Beschäftigungsverhältnis (Wissenschaftlich/ Nicht-Wissenschaftlich)
- Abteilungszugehörigkeit
- Zeitraum der wissenschaftlichen Tätigkeit (Anzugeben sind Jahrzehnte der Beschäftigung)
- Leitungsfunktion
- Freies Textfeld zur evtl. Eingabe personenbezogener Daten

Für die Verarbeitung solcher Daten hatte das Projekt mithin sichergestellt, dass die Datenquelle über hinreichende eigene Berechtigung zur Bereitstellung dieser Daten verfügte oder die Erlaubnis der Urheber nachweisen konnte. Die von den Probanden selbst anzugebenden Informationen wurden auf freiwilliger Basis erhoben.

Die erhobenen personenbezogenen Daten wurden für die Umfrage dem Unterauftragnehmer Görgen & Köller GmbH als verarbeitende Zwischeninstanz auf elektronischem Wege zum Zwecke der Auswertung übermittelt. Zur rechtlichen Absicherung wurde mit dem Unterauftragnehmer ein Vertrag zur Auftragsdatenverarbeitung geschlossen. Der Vertrag sah vor, dass der Auftragsverarbeiter die erhebende Instanz in der Durchführung einer Umfrage zur Tauglichkeit des im Projekt *TextTransfer* entwickelten Impact-Begriffs unter den Mitarbeiterinnen und Mitarbeitern der erhebenden Instanz unterstützte. Bezüglich den Durchführungsmodalitäten waren beide Partner auf die eigens im Projekt entwickelte Umfrageplattform umfragewissen.texttransfer.org beschränkt.[47]

Dem Auftragsverarbeiter war es gestattet, die Umfragedaten für die Dauer der Projektlaufzeit zur Auswertung im Sinne der Fragestellung des Projekts *TextTransfer* zu speichern und zu verarbeiten. Er war verpflichtet, diese Daten spätestens nach 10 Jahren zu löschen. Der Auftragnehmer war nicht berechtigt, diese Daten an Dritte weiterzugeben oder zu veröffentlichen. An der Veröffentlichung der anhand dieser Daten erzielten Analyseergebnisse war der Auftragnehmer beteiligt. Die Befragung war dabei ausschließlich mit Mitarbeiterinnen und Mitarbeitern des Leibniz-Instituts für Deutsche Sprache (IDS) in Mannheim als erhebende Instanz auszuführen. Den Kreis der datenverarbeitenden Instanzen

[47] Die Unterlagen die Datenverarbeitung betreffend sind am IDS gespeichert und können bei Bedarf per Email an bopp@ids-mannheim.de angefordert werden.

grenzte der Vertrag schließlich auf ausschließlich bei Auftraggeber und Auftragnehmer im Projekt *TextTransfer* beschäftigten Mitarbeiterinnen und Mitarbeiter ein.

Eine weitere Aufgabe im AP lag in der qualitativen Aufbereitung der Recherche- und Interviewstudie für tatsächlich erfolgte Verwertung aus dem Pilotprojekt *TextTransfer I* durch den IDS-Projektunterauftragnehmer G&K: Der Aufwand, das Impactpotenzial von Projektberichten zum Großteil analog zu recherchieren, war unter den zur Laufzeit des Projektes herrschenden Strukturen und Vorgaben erheblich. Die Auswertung der in *TextTransfer I* durchgeführten Erhebung zur tatsächlichen Verwertung ließ jedoch einige Ansatzpunkte erkennen, um den Aufwand der Impacterhebung zu reduzieren. So wurde beispielsweise deutlich, dass eine systematisch beibehaltene Impact-Kategorisierung zur zielführenden Nachverfolgung von Impact beiträgt.

Zentrale Erkenntnisse der Aufbereitung konnten insbesondere für die Arbeitsdefinition von „Impact“ sowie die Ausarbeitung des Vorschlags des Kategorienschemas zur Erfassung von Impact in *TextTransfer II* genutzt werden. Außerdem flossen die Ergebnisse der Auswertung in die Instrumente zum Nachweis des tatsächlichen Impacts von *TextTransfer II* ein. Die Studie wurde in der Open-Access-Zeitschriftenreihe des Projektpartners IDS, IDSopen[48], veröffentlicht.

Aufgabe des APs war es außerdem, Strukturen und Wege aufzuzeigen, um eine komfortable Basis zur Nutzung der Methode *TextTransfer* zu schaffen. Für den Bereich der Aufwandsreduzierung in Hinblick auf den Nachweis des tatsächlich erfolgten Impacts, der als Referenz im Rahmen des Maschinellen Lernens von kritischer Bedeutung ist (vgl. Details hierzu Abschlussbericht zum Pilotprojekt *TextTransfer I – AP 2 Stichprobe*), konnte eine Möglichkeit in Form der Adaption des Online-Umfrage-Tools (umfragewissen.texttransfer.org) eruiert werden, wodurch der Auswertungsprozess und der Zugang zu Impactangaben deutlich erleichtert werden konnte.

AP5 war zu Projektende planmäßig abgeschlossen.

[48] Fiedler, N., Köller, Ch., Bopp, J., Schneider, F. (2024): Linguistisches Impact-Assessment: Maschinelle Prognose mit Realitätsabgleich im Projekt TextTransfer. (IDSopen 7). Mannheim: IDS-Verlag.

3.1.6. AP6: Implementierungskonzept

3.1.6.1. Community-gestützte/Offene Bereitstellung und Weiterentwicklung

Die Projektpartner schlagen zur Verstetigung eine offene Implementierung vor. Nicht nur kommt dies der Tatsache entgegen, dass *TextTransfer* mit öffentlichen Geldern entwickelt wurde, die meisten Forschungseinrichtungen kaum Kapazitäten für eine kommerzielle Lösung haben und eine barrierefreie Bereitstellung von digitalen Ressourcen zwischenzeitlich zum guten Standard in der Wissenschaft gehört, verspricht eine Community-gestützte Implementierung das größere Potenzial an Disseminierung, nutzergerechter Anpassung und innovativer Weiterentwicklung im Vergleich zu einer institutionellen Insellösung. Im Rahmen dieses Ansatzes wäre eine Plattform zu wählen, im Rahmen dessen Quellcode und Dokumentation der Methode offengelegt würden, deren Nutzer- und Entwicklercommunity gleichzeitig einschlägige Verbindungen in die Wissenschaft vorzeigen kann.

Beim Umgang mit maschinengestützter Analyse wird im Wissenschaftsbereich zunehmend erwartet, dass die verwendeten Methoden offen und reproduzierbar sind. Das zieht in Bezug auf Large Language Models (LLM) nach sich, dass zugrundeliegende Trainingsdaten (soweit rechtlich und ethisch möglich) zugänglich und dokumentiert sein müssen. Während heute viele in der Industrie verbreitete Sprachmodellen (z. B. Metas LLaMA) selbst bereits frei zur Verfügung gestellt werden, ist die Transparenz und Dokumentation von Trainingsdaten zu Zeiten der Berichtverfassung keineswegs gegeben. Dessen ungeachtet werden Anwendungen auf Basis solcher opaker Sprachmodelle entwickelt und eingesetzt. Sie zu befragen, um Texte aus Wissenschaft und Technik analysieren zu können, ist jedoch problematisch. Es ist bekannt, dass die Ergebnisse dieser Sprachmodelle Fehler und Biases reproduzieren, die oft bereits in ihrem unbekannten Trainingsmaterial enthalten sind. Zudem existiert kein anwendungsbezogener Benchmark, kein Abgleichen oder gar Weiterentwickeln der Modelle auf Grundlage systematisch und transparent erhobener Daten aus der echten Welt, insbesondere nicht für besonders komplexe und spezifischer Fragen wie z. B. jener nach der Verwertbarkeit der Forschungsergebnisse, die in einem Forschungsbericht dokumentiert werden.

In *TextTransfer* wurde daher ein Modell entwickelt, das derartige Fragen mit hoher Validität beantwortet - basierend auf Real-World-Daten. Mit der Lösung der TIB wurde das Modell unter Open-

Source-Bedingungen zur Verfügung gestellt, u.a. auf der Plattform HuggingFace[49], auf deren Playground das Modell von jedermann einfach ausprobiert werden kann. Hier haben interessierte Entwicklerinnen und Entwickler die Möglichkeit, Unzulänglichkeiten dieses Modells zu entdecken, das Modell zu variieren und weiterzuentwickeln. Für Forschungseinrichtungen, Forschungsfördernde, Forschende aus Feldern wie den Science and Technology Studies (STS) und anderen Interessierten werden hier zudem maßgeschneiderte Erweiterungen des Basismodells angeboten, um dabei zu helfen, forschungsfeld- und institutionenspezifische Fragen zu beantworten. Hinzu kommen Lösungsansätze, die (institutionen-intern oder öffentlich) das kontinuierliche Monitoring von sich dynamisch entwickelnden Feldern durch Dashboards und Alert-Dienste unterstützen.

Begleitend zu den oben genannten Kern-Dienstleistungen von *TextTransfer* werden kontinuierlich die Entwicklung vergleichbarer oder verwandter Sprachmodelle seitens Dritter verfolgt, ein öffentlicher Überblick zur Entwicklung dieses Feldes wird geboten, kommentiert und eingeordnet. Zusätzlich findet eine Beratung und Begleitung der Anwendung von Sprachmodellen in der Forschungsbewertung und -Policy unter den Paradigmen von Open Science und FAIRen Forschungsdaten statt.

3.1.7. AP7: Kommunikationskonzept

Eine der wesentlichen Aufgaben des AP 7 „Kommunikationskonzept“ bestand darin, Wege zu entwickeln, um die Arbeiten am Projekt wie auch das Projektergebnis in den anvisierten Zielgruppen bekanntzumachen.

3.1.7.1. Webpräsenz texttransfer.org

Entsprechend der Vorhabenbeschreibung wurde im Rahmen des Kommunikationskonzeptes unter der Federführung des IDS von beiden Partnern eine eigene Webseite für das Projekt aufgebaut, die unter der Domain *texttransfer.org* registriert ist.

[49] https://huggingface.co/

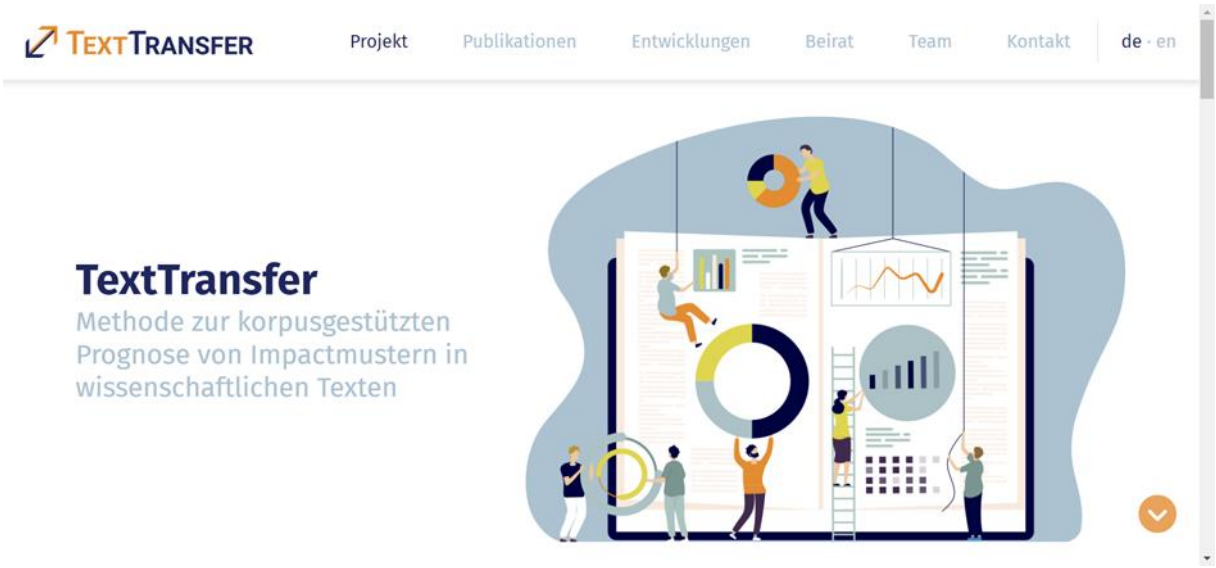

Abbildung 3-15 Die Web-Präsenz des Projektes unter „texttransfer.org“.

Die Web-Präsenz diente dazu, die Sichtbarkeit des Projektes zu erhöhen, indem sie u. a. den Ansatz als gefördertes Vorhaben dokumentierte, die unterschiedlichen Player und ihre Rollen vorstellte, als auch für Interessierte an der Methode eine Kontaktmöglichkeit für weitergehenden Austausch in einfacher Weise zu ermöglichen. Die Web-Seite wurde in deutscher und englischer Sprache angelegt, um die Reichweite zu erhöhen.

Zwecks Konzeption der Website wurde eine spezielle Taskforce zum Thema „Reformulierung Mission *TextTransfer* für die neue Website“ aus Mitarbeiterinnen und Mitarbeitern der Projektpartner TIB und IDS gebildet, die sich mit den Inhalten und dem Design des neu zu gestaltenden Webauftritts befasste. Ziel war es, die Website – sowohl was die Struktur als auch die Inhalte betraf – so zu gestalten, dass die Idee und das Anliegen von *TextTransfer* auch Besucherinnen und Besuchern, die wenig bis kein Hintergrundwissen zum Thema hatten, so verständlich wie möglich vermittelt wurden und somit mögliche Barrieren, sich mit dem Lösungsansatz auseinanderzusetzen, möglichst gering zu halten.

Unter diesem Aspekt konnte für die Formulierung und Erstellung der kritischen textuellen Inhalte – nach einem entsprechenden Briefing durch die IDS-Kolleginnen und -Kollegen – ein Mitarbeiter des

Projektpartners TIB gewonnen werden, der sich schwerpunktmäßig mit dem Thema Wissenschafts-kommunikation beschäftigt.

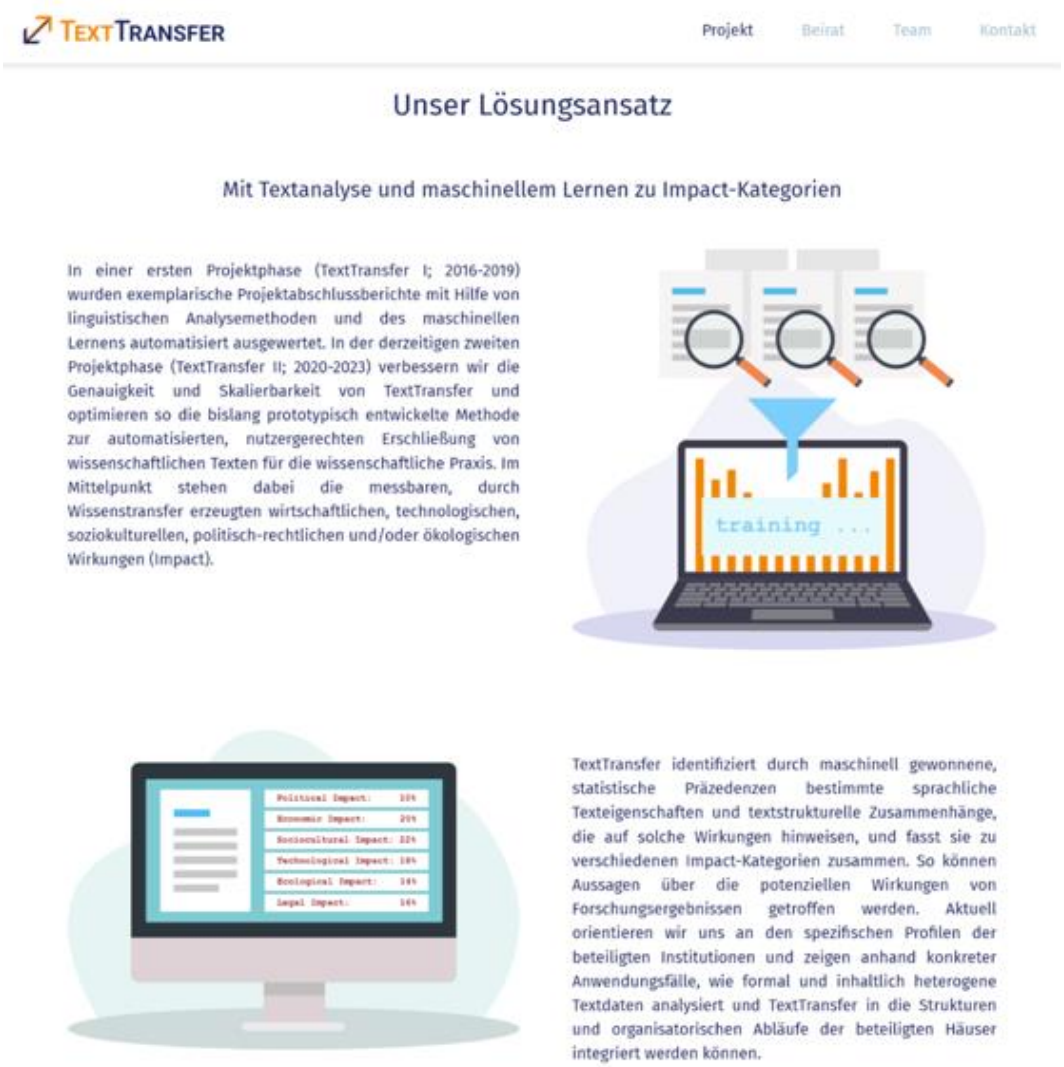

Abbildung 3-16 Unter „texttransfer.org“ wird der Ansatz der Methode TextTransfer leicht verständlich erklärt.

Der Projektpartner IDS kümmerte sich um die Erstellung der grafischen Inhalte, das Gesamtlayout sowie die technische Umsetzung der Website.

Die Website wurde in HTML, CSS und JavaScript implementiert. Zusätzlich wurde das CSS-Framework Bootstrap verwendet. Für die Funktion des Kontaktformulars wurde auf die Library PHPMailer, welche auf der Programmiersprache PHP basiert, zurückgegriffen.[50] Die Domain „texttransfer.org“ sowie ein Webhosting-Paket zur Veröffentlichung der Website wurden bei lima-city.de – finanziert durch IDS-Haushaltsmitteln – angemietet.

3.1.7.2. Beirat

Ein weiterer Meilenstein bildete im Rahmen von AP 7 die Bildung eines Beirats, der das Projekt sowohl in wissenschaftlichen als auch zu weiteren Themen, u.a. der Dissemination, unterstützen sollte.

Der Projektbeirat wurde im ersten Projektjahr durch Mitarbeiter der Partner selbst eingerichtet:

- Prof. Dr. Ina Blümel[51], Stellvertretende Leiterin Open Science Lab des Projektpartners TIB, die gleichzeitig auch Dozentin für Informationsmanagement an der Hochschule Hannover ist
- Dr. Walter Görgen[52], Geschäftsführer des Unterauftragnehmers G&K
- Dr. Marc Kupietz[53], Leiter des Programmbereichs Korpuslinguistik des IDS

Die genannten Personen waren zur fachlichen Begleitung des Vorhabens bestellt.

Im zweiten Projektjahr konnten fünf weitere, projekt-externe Beiräte gewonnen werden, die als Multiplikatorinnen und Mulitplikatoren bzw. „TransferBeiräte“ fungierten:

- Lindsay Green Barber[54], Founder and CEO of Impact Architects, Expertise für qualitative Methoden des Impact Assessment
- Cornels Lehmann-Brauns[55], Stifterverband, Projektleiter des Evaluierungsinstruments "Transferbarometer"

[50] Auf Grund des Umfangs der Skripte, sind diese nicht Teil des Berichts, können jedoch auf Anfrage per Mail an bopp@ids-mannheim.de zur Verfügung gestellt werden.
[51] https://vivo.tib.eu/fis/display/n0000-0002-3075-7640
[52] https://gk-mb.com/unser-unternehmen/team
[53] https://www.ids-mannheim.de/digspra/personal/kupietz/
[54] https://www.theimpactarchitects.com
[55] https://www.stifterverband.org/transferbarometer

- Dr. Mark Mann[56], Innovationsberater und Projektmanager – insbesondere für den Bereich der Anwendbarkeit geistes- und sozialwissenschaftlicher Forschung
- Dr. Florian Schütz[57], Geschäftsführer der KI Park e.V., Berlin
- Christine Wennrich[58], Leiterin des Dezernats Transfer der Leibniz-Gemeinschaft, Beiratsmitglied im Partnervorhaben ExpResViP an der TIB

Kriterium für die Auswahl eines Kandidaten/einer Kandidatin war dabei ein technischer Hintergrund, der die Person jeweils in die Lage versetzen sollte, sowohl die Projektergebnisse zu bewerten als auch Statements hierzu abzugeben bzw. als Multiplikator/Multiplikatorin zu fungieren.

Gemeinsam mit dem fachlichen Beirat des Projekts gab es im März 2023 ein Projekttreffen beim Verbundpartner TIB in Hannover, bei dem die Beiräte Rückmeldungen zu den zu diesem Zeitpunkt bestehenden Projektergebnissen als auch Anregungen zu weiteren Entwicklungen bzw. zur Dissemination gaben. Insgesamt wurden das Projekt und seine Ergebnisse bis dato als sehr positiv von dem Fachgremium aufgenommen, die Methode als ausgesprochen mächtig anerkannt. Zentrale Themen waren ethische und rechtliche Implikationen von *TextTransfer*, wie sie sich insbesondere aus einer potenziell missbräuchlichen Nutzung (z. B. Retro-Engineering) der Methode ergeben könnten.

Mit dem Abschlusstreffen Ende Mai 2024 beim Verbundpartner IDS in Mannheim, ebenfalls unter Beteiligung des Beirats, konnte das Projekt erfolgreich zum Abschluss gebracht werden.

AP7 war zu Projektende planmäßig abgeschlossen.

3.1.8. AP8: Projektmanagement

Als gesamtverantwortliche Institution hatte das IDS im Verbund sowohl die Projektleitung als auch das Gesamt-Projektmanagement inne und koordinierte somit 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.

[56] https://www.markmann.ee
[57] https://kipark.de/team-der-geschaeftsstelle/
[58] https://www.leibniz-gemeinschaft.de/transfer/wissens-und-technologietransfer

Zu Projektbeginn hatten daher alle Projektpartner entsprechende interne Maßnahmen getroffen, die einen reibungslosen Ablauf des Projektes sicherstellten (u.a. Einrichtung einer internen Infrastruktur mit beispielsweise projektspezifischem Exchange-Server, Zuständigkeiten etc.).

Als zentrale Kommunikationsstruktur, insbesondere zum Dokumentenaustausch der Partner, diente – neben Emailanwendungen – ein virtueller Projektordner innerhalb einer IDS-internen Cloud-Lösung, der eigens für das Projekt erstellt wurde und vom IDS administriert und organisiert wird. Je nach Rolle im Projekt bzw. nach Zusammenarbeit in den verschiedenen APs hatten, auch IDS-externe Projektmitarbeiterinnen und Projektmitarbeiter entsprechende Zugriffs- und Arbeitsrechte. Der Projektstand als auch Vor- und Nachbereitungen zu Meetings wurden über ein Rolling Document kommuniziert, zu dem alle Projektmitarbeitende Zugriff hatten. Damit war stets eine Transparenz zu den einzelnen Aktivitäten im Projekt gewährleistet. Auch Team-Mitglieder, die nicht aktiv an einem AP beteiligt waren, könnten sich somit aktiv jederzeit über den Stand der Arbeiten im Projekt informieren.

Sonstige Workflows und Pipelines wurden gemäß Projektfortlauf eingerichtet.

Für den Austausch sehr großer Datenmengen zwischen einzelnen Partnern wird u.a. auch die Anwendung Gigamove der RWTH Aachen bzw. des DFN genutzt. Kollaboratives Arbeiten an Dokumenten erfolgte außerdem über Google Docs und den LaTeX Editor Overleaf.

Für einen persönlichen Austausch – außerhalb von physischen Treffen – innerhalb des Projektes wurde neben der Telefonie auf die Videokonferenzsysteme des vom Projektpartner IDS angebotenen Tools Zoom zurückgegriffen.

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

AP8 war zu Projektende planmäßig abgeschlossen.

3.2. Die wichtigsten Positionen des zahlenmäßigen Nachweises

Das Projekt wurde zum 26.06.2020 rückwirkend zum 01.06.2020 bewilligt. Pandemiebedingt unterlagen beide Projektpartner seit März 2020 starken Betriebseinschränkungen bis zur Schließung.

Aufgrund des unter Bedingungen der Corona-Pandemie erfolgten Projektstarts ergaben sich für den Projektpartner IDS mit Blick auf die Kostenplanungen Änderungen der ursprünglichen Antragsplanung. In der Folge konnten die im Projekt vorgesehenen Mitarbeiterstellen zunächst nicht gemäß personellen und sozialen Qualitätsstandards ausgeschrieben und plangemäß besetzt bzw. Kooperations- und Forschungsverträge nicht zeitnah unterzeichnet werden. Die Rahmenbedingungen am Arbeitsmarkt sowie die starken Mobilitätseinschränkungen für wissenschaftliche Mitarbeiter in Zeiten der Corona-Pandemie hatten überdies dafür gesorgt, dass der wissenschaftliche Personalbestand des Vorhabens unvorhersehbar eingebrochen war. Die freigewordenen Mittel wurden daher in Abstimmung mit dem Projektträger zum Dezember 2020 in das Jahr 2021 übertragen und eingeplant.

Beide für die wissenschaftlichen Stellen im Projekt vorgesehenen Mitarbeiter hatten aus im familiären Umfeld der Betroffenen angesiedelten Faktoren außerdem ihr Arbeitsverhältnis ganz oder langfristig beendet. In der Folge war dem Vorhaben ein erheblicher Minderbedarf entstanden. Dem IDS war es nach intensiven Bemühungen durch Neuausschreibungen und umfassenden Qualifizierungsmaßnahmen von wissenschaftlichen Nachwuchskräften gelungen, die zur Zielerreichung im Projekt notwendigen Kapazitäten wiederaufzubauen. Um die bewilligten Mittel sachgerecht einsetzen zu können, hatte das Teilprojekt in Abstimmung mit dem Projektträger eine Verschiebung in der Ausgabenplanung beantragt, die zum 07.10.2021 genehmigt wurde.

Dank umfassender personeller Umstrukturierungen konnte letztendlich der gewünschte Arbeitsstand weitgehend in 2022 wiederhergestellt werden. Die elternzeitbedingte Vertretungsregelung zur Kompensation des Wegfalls einer wissenschaftlichen Stelle im Projekt lief zum 31.12.2022 aus, so dass die vorgesehene Mitarbeiterin dem Projekt bis zu seinem regulären Ende zur Verfügung stehen konnte. In der Folge wurden in enger Abstimmung mit dem Projektträger infolge einer geänderten kassenmäßigen Bereitstellung der Projektmittel vom 07.10.2021 eine Neuverteilung der Arbeitslast im Vorhaben geplant und umgesetzt. Das IDS-Teilprojekt hatte daher am 05.10.2022 eine Verschiebung geringer

Restmittel in das Haushaltsjahr 2023 beantragt, dem mit einem Schreiben vom 21.11.2022 durch den Projektträger stattgegeben wurde.

Projektträger und Projektleitung des IDS kamen darin überein, dass die planmäßige Durchführung des beantragten Ansatzes erfüllt werden könnten, sofern *TextTransfer* die notwendigen zeitlichen Reserven zur Kompensation seiner Ausfälle eingeräumt würde. Im Zuge der Verzögerungen und insbesondere zur Absicherung der Projektziele wurde in 2023 – wie bereits in 2022 vorgesehen – eine kostenneutrale Verlängerung für den Zeitraum eines Jahres durch ein Anschreiben vom 19.04.2023 durch das IDS beantragt und durch den Zuwendungsgeber DLR mit dem Bescheid vom 17.05.2023 genehmigt. Die dazu erforderlichen Mittel standen dem Projekt nach modifiziertem Abrufungsturnus zur Verfügung. Der Arbeits- und Zeitplan wurde entsprechend ausgeführt, so dass das Gesamtprojekt zum 31.05.2024 ordnungsgemäß abgeschlossen werden konnte. Alle Partner, neben der TIB auch die Unterauftragnehmer G&K und UIUC, waren dabei kostenneutral in die Verlängerung mitgegangen.

Sämtliche Auszahlungsanordnungen für die Unterauftragnehmer G&K und UIUC des Projektpartners IDS für den ursprünglich geplanten Projektzeitraum, d.h. 31.05.2023, wurden ordnungsgemäß abgerufen.

Dem Projektpartner TIB gelang es, seine Arbeiten im Projekt in dem ursprünglich geplanten Zeitraum bis zum 31.05.2023 zu beenden. Das Budget der TIB wurde zu 100 % für die Personalkosten verwendet. Dienstreisen haben aufgrund der Corona-Maßnahmen im Projektzeit nicht stattgefunden, das beantragten Budget wurde daher nicht genutzt. Die Mittel wurden stattdessen zur Deckung der erhöhten Personalkosten genutzt.

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 die Entwicklung und Bereitstellung 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 einem bedarfsgerecht zugeschnittenen und konvertierten Korpus von Forschungsberichten als Stichprobe nötig.

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

FKZ 01IO2002A
Version 1.1

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

Das Projekt *TextTransfer* 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 erbracht. Das Hauptprojekt setzte in einer zweiten Förderphase darauf auf und versuchte die Methode durch die Einbeziehung größerer Lern- und Metadatenbestände und zusätzlicher Lernansätze in ihrer Leistungsfähigkeit zu skalieren und in ihrer Präzision zu stabilisieren. Der Quelltyp Projektabschlussbericht wurde dabei als geeignete Grundlage maschineller Prognosen etabliert, die Prozesse in der Methodenanwendung durch Automatisierung in ihrer ökonomischen Effizienz optimiert.

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 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 stellte 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 befand. Während kommerzielle Analyseverfahren, die zumeist auf englischsprachigen Daten basieren, an Zahl und Reichweite gewinnen, stieß *TextTransfer* explizit in die Nische bisher unerschlossener deutschsprachiger Daten. Auf diese Weise stellte sich die Methode nicht nur hinsichtlich ihres Fähigkeitsprofils, die Wahrscheinlichkeiten von Impactpotenzial von Forschungsergebnissen aufgrund statistischer Vergleiche zu prognostizieren, als

Innovation auf. Sie leistete außerdem ihren Beitrag in der Nutzung künstlicher Intelligenz in Deutschland und öffnet nationale Kernmärkte für neue Ansätze des Impact Assessments.

Die Methode *TextTransfer* kann somit zu einem Instrument des Wissens- und Technologietransfers werden, da es den beteiligten Instituten auch in Zusammenarbeit mit wissenschaftlichen Dachorganisationen wie der Leibniz-Gemeinschaft erlaubt, eigene Projektberichte 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, die im Rahmen des Lerndatensatzes in der Vergangenheit spezifische Impacterfolge nach sich gezogen haben. Entsprechend ist das künstlich über die Methodenanwendung generieret Erfahrungswissen in der Impactbewertung von wissenschaftlichen Einrichtungen statistisch einzuschätzen.

Sämtliche Ergebnisse des Forschungsprojekts (Quellcode, Dokumentation) werden quelloffen zur Verfügung gestellt und können von potenziellen Anwendern künftig frei genutzt werden. Der Mehrwert der Methode *TextTransfer* gegenüber agenturgestützten kommerziellen Lösungen ist primär darin zu sehen, dass letztere hinter einer Bezahlschranke nicht barrierefrei genutzt werden können. Mit Blick auf die Methodenmächtigkeit ist *TextTransfer* in der Anwendung auf breite, gesamtgesellschaftliche Anwendungen über rein kommerzielle Impactszenarien hinaus bewährt. Auch sein Fokus auf deutschsprachige Daten füllt ein Desiderat in der Anwendung KI-gestützter Verfahren des Text Mining.

Darüber hinaus wurden Vorschläge erarbeitet, wie technische und rechtliche Rahmenbedingungen etabliert werden können, in Berichtsform publizierte Forschungsergebnisse bald einer zeitgemäßen Nutzung und Analyse zu unterziehen. Die Projektbeteiligen appellieren an die Fördermittelgeber perspektivisch entsprechende Standards für die Gewährleistung von Maschinenlesbarkeit einzufordern.

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 der Methode *TextTransfer* 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 *TextTransfer* bisher unbearbeitetes Terrain dar. Komplementäre rein kommerzielle Entwicklungen zielen auf einen ähnlichen Ansatz mit einem sehr viel stärker monetär eingrenzten Impactbegriff, die Auswertung verläuft entlang fokussierter Textabschnitte, industrieller Schlüsselbegriffe und Abgleich mit bestehenden Patenten.[59] Das Alleinstellungsmerkmal von *TextTransfer* ergibt sich sowohl mit Blick auf die Quelldaten, als auch auf die Metadaten aus der Analysefähigkeit gesamtgesellschaftlicher Anwendungsfelder.

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

3.6.1. Vorträge

- ANDREAS WITT (IDS)/MARIA BECKER (IDS): Vortrag “TextTransfer: Methods for corpus-based prediction of impact in scientific texts". Lecture series Data Science in Action, Mannheim, 27.10.2022
- ANDREAS WITT (IDS)/MARIA BECKER (IDS): Posterpräsentation „TextTransfer - Methode zur korpusgestützten Prognose von Impactmustern in wissenschaftlichen Texten.“, DLR 12. Transferwerkstatt, Berlin, 17./18.11.2022.

[59] https://scoutinscience.com/faq/how-does-the-algorithm-work

- ANDREAS WITT (IDS): Vortrag „TextTransfer – Methods of impact assessment on the basis of machine readable project final reports.", ASTP The Monthly online SSHA impact flashlight, 07.12.2022.
- MARIA BECKER (IDS): Vortrag "Methoden zur korpusbasierten Wirkungsvorhersage in wissenschaftlichen Texten". Kolloquium des Deutschen Zentrums Für Hochschul- und Wissenschaftsforschung (DZHW), Hannover, 23.05.2023.
- MARIA BECKER (IDS): Vortrag "Automatisierte Prognose und Klassifikation des Impacts wissenschaftlicher Forschung innerhalb der Wissenschaft und darüber hinaus." Ringvorlesung „Aktuelle Fragen der Sprach- und Translationswissenschaft", Institut für Übersetzen und Dolmetschen Heidelberg, 18.01.2024
- MARIA BECKER (IDS): Automatisierte Analyse des gesellschaftlichen Impacts von Forschung mit Sprachmodellen. 3. Text+ Plenary 2024, Mannheim, 10.10.2024.

3.6.2. Veranstaltung, Workshops, Kurse

- 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.[60]
- JANA DIESNER (UIUC), REZVANEH REZAPOUR (UIUC), MARIA BECKER (IDS): Session "Impact assessment and network analysis", Sunbelt 2022, online, 13.07.2022.

3.6.3. Publikationen/Poster

- LAMBERT HELLER Proposal Pilot zur Erstellung von strukturierten Forschungsberichten https://proposalpilot.texttransfer.org/document.php?name=TextTransfer%20Pilot.json (2023)
- MARIA BECKER (IDS), KANYAO HAN, ANTONINA WERTHMANN (IDS), SHADI REZAPOUR, HAEJIN LEE, JANA DIESNER (UIUC), ANDREAS WITT (IDS): Detecting Impact Relevant Sections in Scientific Research. In: Proceedings of LREC. Torino, Italy (2024).

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

- NORMAN FIEDLER (IDS), CHRISTOPH KÖLLNER Köller, JUTTA BOPP (IDS), FELIX SCHNEIDER: Linguistisches Impact-Assessment: Maschinelle Prognose mit Realitätsabgleich im Projekt TextTransfer. (= IDSopen 7). Mannheim: IDS-Verlag (2024).
- NORMAN FIEDLER (IDS): Die Sprache des Impacts. Wie ausgerechnet die Linguistik Wirkung und Anwendung von Forschung vorhersagt. (= Transfer & Innovation 2024-2), S. 109-126. Berlin: DUZ Verlag (2024)
- LAMBERT HELLER: Open Source Quellcode Erstellung Projektberichte https://github.com/TIBHannover/text-transfer-ii-prototype (2024)
- MARIA BECKER (IDS), KANYAO HAN, SHADI REZAPOUR, JANA DIESNER (UIUC), ANDREAS WITT (IDS): Impact Classification within and beyond Academia: Domain-Robust Annotation and the Capacity of Large Language Models[61]
- MARIA BECKER (IDS), MATTHES FÜRST, ANDREAS WITT (IDS): Societal Impact of Scientific Research: A corpus linguistic analysis of media resources[62]

_____ _____

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

[61] Zum Zeitpunkt der Berichtsverfassung noch in Vorbereitung.
[62] Zum Zeitpunkt der Berichtsverfassung noch in Vorbereitung.

FKZ 01IO2002A
Version 1.1

4. Anlagen

4.1. Ad AP 2: Anwendung Online-Umfrage-Tool „umfragewissen.texttranfer.org“

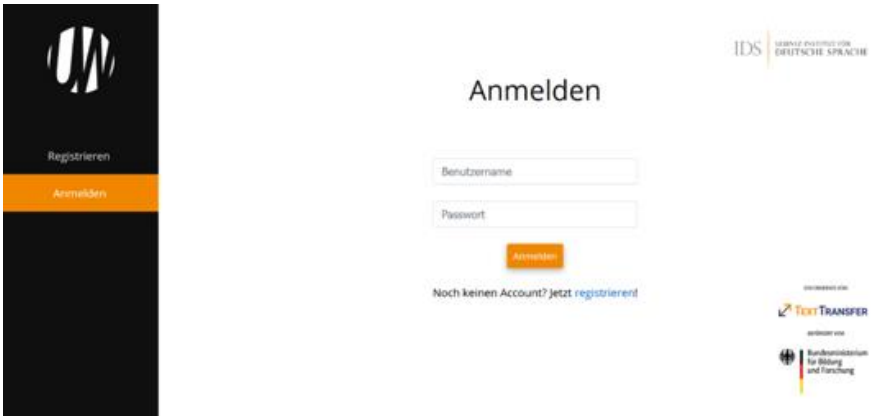

Abbildung 4-1 Anmeldeseite des Online-Umfrage-Tools „umfragewissen.texttransfer.org“

Um *umfragewissen.texttransfer.org* nutzen zu können, ist die Registrierung eines Nutzer-Accounts erforderlich.

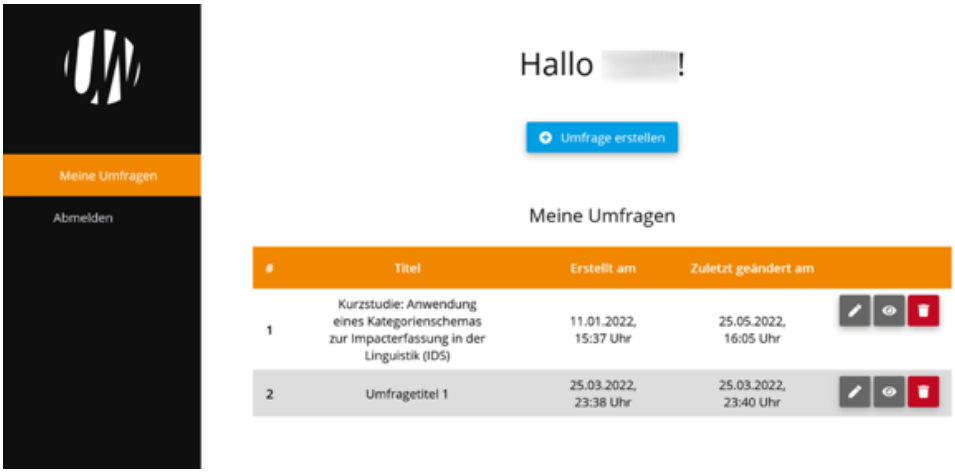

Abbildung 4-2 Übersichtsseite des angemeldeten Nutzers in „umfragewissen.texttransfer.org“

Nach erfolgreicher Registrierung und Anmeldung erfolgt eine Weiterleitung auf die persönliche Übersichtsseite, auf der sowohl neue Umfragen angelegt als auch bereits erstellte Umfragen überarbeitet, angesehen und gelöscht werden können. Bei Erstellung einer neuen Umfrage sowie bei Überarbeitung eines bereits erstellten Umfrageformulars werden jedes Mal automatisch Datum und Uhrzeit erfasst und angezeigt.

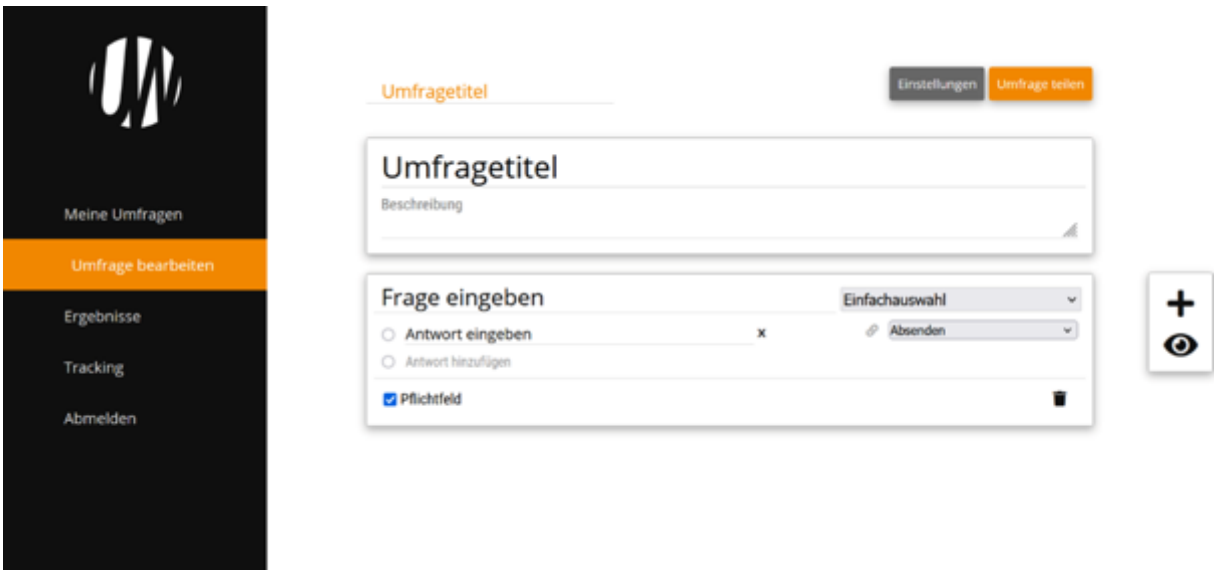

Abbildung 4-3 Erstellung bzw. Überarbeitung einer Umfrage mittels „umfragewissen.texttransfer-org“

Das Tool bietet eine schnelle und unkomplizierte Möglichkeit zur Erstellung von neuen Umfragen, ohne eine große Einarbeitung vorauszusetzen. Es können neben Multiple-Choice-Fragen ebenso Fragen mit Freitext-Antworten sowie Skalen verwendet werden. Zudem kann bei einer Multiple-Choice-Frage mit nur einer Antwortauswahlmöglichkeit angegeben werden, dass Teilnehmende – je nach Antwort – zu einer bestimmten nächsten Frage weitergeleitet werden sollen.

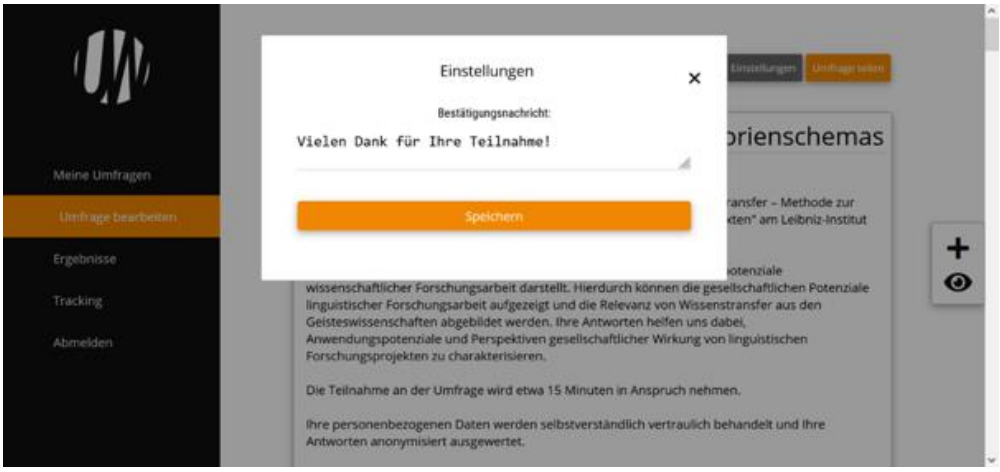

Abbildung 4-4 Einstellungsmöglichkeit einer individuellen Teilnahmebestätigungsnachricht für Umfrageteilnehmende

Nutzerinnen und Nutzer können unter dem Menüpunkt „Einstellungen“ außerdem eine individuelle Teilnahmebestätigungsnachricht für die Probandinnen und Probanden verfassen. Auf diese Weise können beispielsweise auch weiterführende Hinweise zur Studie hinterlegt werden.

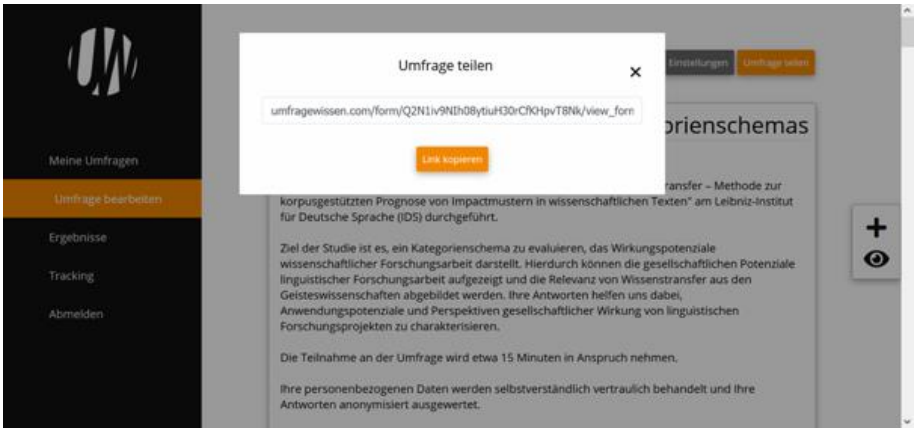

Abbildung 4-5 Automatische Link-Generierung zum Versenden der Umfrage an potenzielle Teilnehmende

Nach Fertigstellung einer Umfrage kann diese mit Hilfe eines automatisch generierten Links, welcher unter dem Button „Umfrage teilen“ zu finden ist, an potenzielle Probandinnen und Probanden verschickt werden.

Abbildung 4-6 Ansicht für Teilnehmende einer Umfrage in „umfragewissen.texttransfer.org“

Teilnehmende, die den Einladungslink zur Umfrage anklicken, werden entsprechend auf die Umfrageseite weitergeleitet.

FKZ 01IO2002A
Version 1.1

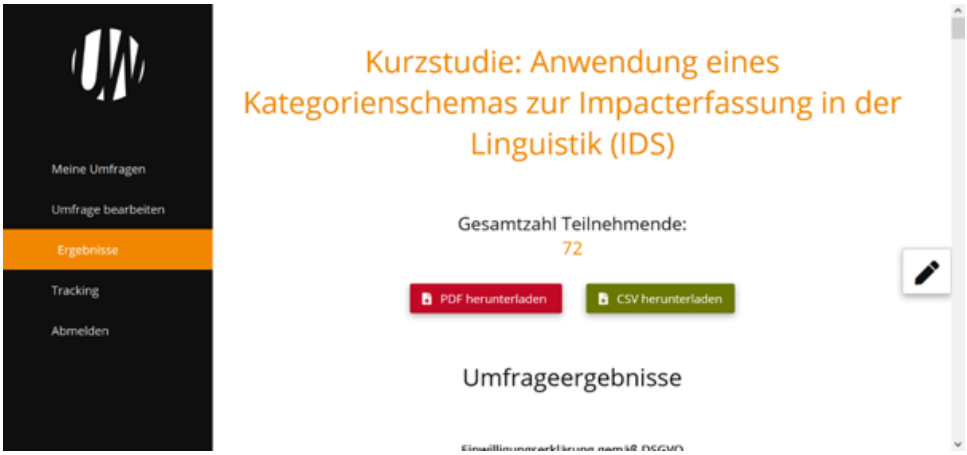

Abbildung 4-7 Übersicht und Download-Möglichkeit der Ergebnisse in „umfragewissen.texttransfer.org“

Die Umfrageergebnisse können jederzeit im Bereich „Ergebnisse“ einer Umfrage eingesehen werden. Hier können diese dann sowohl als PDF- als auch als CSV-Datei heruntergeladen werden. Insbesondere der Download im CSV-Format bietet eine schnelle Auswertungsmöglichkeit in Anwendungen wie *Excel* oder anderen Analyse- und Visualisierungstools.

Abbildung 4-8 Tracking einzelner Teilnehmerinnen und Teilnehmer in „umfragewissen.texttransfer.org“

FKZ 01IO2002A
Version 1.1

Mit Hilfe der Tracking-Funktion können zum einen die einzelnen Antworten der Teilnehmenden genauer betrachtet und zum anderen die Zeit, welche die Teilnehmenden zur Beantwortung der Umfrage benötigt haben, eingesehen werden.

FKZ 01IO2002A
Version 1.1

4.2. Ad AP 2: Codebook

Annotation von Impact-Kategorien

ANNOTATIONSSCHEMA

[Deutsche Version]

Letztes Update: 29.11.2022

Bitte lesen Sie das Annotationsschema sorgfältig durch. Machen Sie sich mit den Impact-Kategorien, den Erklärungen, Definitionen und Beispielen vertraut, bevor Sie mit dem Annotieren beginnen.

Ziel der Annotationen

Ziel dieses Annotationsexperiments ist es, textbasierte Wirkungsnachweise in wissenschaftlichen Projektberichten zu identifizieren und zu kategorisieren – also alle Auswirkungen, die ein Forschungsprojekt auf das eigene wissenschaftliche Feld, auf die Gesellschaft, die Umwelt usw. hat oder haben kann.

Wir definieren Impact als Wirkung wissenschaftlicher Aktivitäten innerhalb der Wissenschaft oder außerhalb des akademischen Bereichs, z. aus Wissenschaft, Wirtschaft, Gesellschaft, Kultur, Politik, Recht, Technik oder Umwelt.

In einem Bericht kann die Wirkung durch die Beschreibung von Methoden und Routinen dargestellt werden, die für ein Projekt implementiert wurden, oder durch die Wirkung, die die Autoren beim Schreiben ihrer Abschlussberichte erwartet haben (→ estimierte Auswirkungen werden auch als Impact betrachtet!).

Die Projektberichte stammen aus vier verschiedenen Bereichen: KI, Linguistik, Elektromobilität und Musikwissenschaft. Der Impact kann domänenspezifisch (betrifft die jeweilige Domäne) oder sehr allgemein sein – z. beeinflusst die akademische Gemeinschaft, die Gesellschaft usw.)

Annotationstool

Wir verwenden die browserbasierte Version von Inception als Annotationstool, stellen Sie also sicher, dass Sie beim Annotieren eine stabile Internetverbindung haben. Es ist keine Installation erforderlich. In Inception stellen wir Ihnen die Texte, die annotiert werden, sowie das Labelset mit den Impact-Kategorien zur Verfügung.

Labelset (Kategorienschema)

FKZ 01IO2002A
Version 1.1

Beim Annotieren werden Ihnen Passagen zur Verfügung gestellt, die wahrscheinlich auf Auswirkungen hinweisen. Wählen Sie für die Sätze, die Auswirkungen anzeigen, die am besten geeignete Haupt- und Unterkategorie aus, wie unten in der Tabelle beschrieben. Wenn zwei Kategorien gleichermaßen geeignet sind, können Sie mehr als ein Label auswählen (sowohl aus Haupt- als auch aus Unterkategorien).

Wenn Sie sich bei der Hauptkategorie sicher sind, aber keine der Unterkategorien passt, wählen Sie die jeweilige Hauptkategorie und als Unterkategorie „**Other**". Bitte fügen Sie einen Vorschlag für ein passendes Unterlabel ein, indem Sie ihn in das Feld „**Label Suggestion**" eingeben.

Wenn Sie sicher sind, dass Impact vorliegt, aber keine der Hauptkategorien passt, wählen Sie die Hauptkategorie „**Other (Main)**". Bitte geben Sie einen Vorschlag für ein passendes Etikett ein, indem Sie ihn in das Feld „**Label Suggestion**" eingeben.

Wenn ein Satz keine Wirkung ausdrückt (was der Fall sein könnte!), wählen Sie die Bezeichnung „**No Impact**". In diesem Fall wird keine der anderen Kategorien annotiert.

Wir annotieren auf **Satzebene**. Bitte markieren Sie bei der Etikettenauswahl den vollständigen Satz. Falls es zusätzliche Wörter gibt, die nicht zum Satz gehören (z. B. Überschriften am Anfang eines Satzes), werden diese nicht markiert (z. B.: *Zielsetzung Das Projekt verfolgt das Ziel….* → markieren Sie hier nicht das Wort *Zielsetzung*).

Es gibt ein paar zusätzliche Labels, die Sie verwenden können/sollten:

Element Priority: Ist der Satz für die Angabe der Wirkung hoch relevant, oder mittel oder nur gering?

- Bitte annotieren Sie dieses Label für jede Instanz, die Sie auch mit einer der Impactkategorie gelabelt haben.
- Wählen Sie das Label „**HIGH**" für eine starke Wirkung (z. B. Das Produkt hat die Lebensqualität der Studienteilnehmer langfristig verbessert), „**MEDIUM**" für eine „normale" Wirkung (z. B. Durch den Workshop konnten neue Kontakte aufgebaut werden) und „**LOW**" für Grenzfälle (wenn es nur eine kleine Auswirkung gibt oder wenn Sie sich nicht sicher sind)
- Manchmal weisen Signalwörter wie „nachweisliche Auswirkung" etc. auf eine starke Auswirkung hin - sie sind aber für die Wahl des Labels „HIGH" nicht zwingend notwendig.
- Die Frage, ob eine Wirkung zu erwarten oder bereits nachgewiesen ist, spielt für die Wahl der Elements Priorität keine Rolle.

Snippet: Es kann sein, dass die Sätze falsch gesplittet wurden. Wenn Sie zwei (oder mehr) Annotationsinstanzen finden, die einen Satz bilden, weisen Sie das Impact-Label nur der letzten zu und kommentieren Sie die erste (oder mehrere) Instanzen mit dem Label „**Snippet**".

Comment: Sie können dieses Feld verwenden, um Ihre allgemeinen Gedanken aufzuschreiben, z. B. die allgemeinen Herausforderungen beim Annotieren der Daten, wenn Sie sich bei einer Kategorie nicht sicher sind, wenn Sie Fragen oder Beobachtungen haben, usw.

Weitere Hinweise**…**

- Jeder Satz sollte isoliert betrachtet werden, d.h. im Satz selbst müssen konkrete Wirkungsindikatoren enthalten sein. Es reicht nicht aus, wenn der Satz in einem großen Zusammenhang die Wirkung früherer Sätze „bestätigt".
- Wenn zwei oder mehr Sätze eine „Einheit" bilden, also alle Indikatoren für dieselbe Kategorie enthalten und sich gegenseitig verstärken, sollten sie dennoch isoliert annotiert werden (d.h. Sie annotieren die Sätze getrennt mit denselben Kategorien).
- Wichtig: Besprechen Sie Ihre Arbeit nicht mit den anderen Annotatoren und Annotatorinnen oder Personen außerhalb dieses Projekts.

Impact Kategorien

HAUPT-KATEGORIE	Beschreibung und Unterkategorien	Beispiele
Gesellschaft-licher Impact	Auswirkungen auf gesellschaftliche Gruppen oder Institutionen wie Schulen, Kommunen, Stiftungen oder Vereine, Flüchtlinge/Migration etc.	
	Bildung außerhalb der Wissenschaft, z.B. neue/verbesserte Lern- und Lehrmethoden für Schulen, Lerneffizienz, Erlernen praktischer Fähigkeiten etc.	[MUS] Die zeitliche Struktur der Weiterbildung der Gymnasiallehrer wurde trotz der genannten Schwierigkeiten bei der Vereinbarkeit von Familie und Beruf als positiv bewertet, was für den langen Weiterbildungszeitraum von MuBiKi spricht.
	Kultur, z.B. die Organisation von Konzerten, Theatern, Lesungen etc.	[MUS] Am Ende des Workshops gab die Teilnehmerin ein Konzert beim Altstadtfestival, bei dem sich die Kinder für zukünftige Kurse anmelden konnten.
	Körperliche Gesundheit, z.B. weniger Atemwegserkrankungen, Impfkampagnen etc.	[KI] Mit Hilfe der im Projekt entwickelten Trenderkennung auf Basis erfasster Prozesse und Vitalparameter soll es künftig möglich sein, mögliche Gefahren in der Entwicklung des Krankheitsverlaufs frühzeitig zu erkennen und einzuleiten geeignete Maßnahmen in der Versorgung des jeweils betroffenen Patienten.
	Lebensqualität/Psychische Gesundheit, z.B. weniger Depressionen, Work-Life-Balance, Smart-Home-Systeme, persönliches Wachstum etc.	[KI] Die Erkenntnisse aus diesem Projekt sollen zu einer besseren Versorgung hilfsbedürftiger älterer Menschen beitragen.
	Sicherheit, z. B. Verkehrssicherheit, Benutzersicherheit, allgemeine Sicherheit, weniger Unfälle… (jedoch	[MOB] Ein weiteres Ergebnis des Projekts war die Ableitung von Empfehlungen für

FKZ 01IO2002A
Version 1.1

	nicht IT-Sicherheit/Datenschutz; siehe unten)	Sicherheitsstandards für Elektrofahrzeuge auf Basis der Ergebnisse von Nutzerstudien.
	Aufbau/Verbesserung von Kooperationen oder Netzwerken (außerhalb der Wissenschaft)	[MUS] Eines der interessantesten und wichtigsten Ergebnisse ist sicherlich die Zusammenarbeit mit Lehrkräften an Gymnasien, die sich von arbeitsteiligen Formen der Zusammenarbeit abhebt und vor dem Hintergrund einer Erweiterung beruflicher Handlungsspielräume eine gleichberechtigte Zusammenarbeit ermöglicht.
	Sonstige (Other)	
Politischer und rechtlicher Impact	Verwendung der Projektergebnisse in politischen oder gesetzgeberischen Kontexten	
	Politische Regulierungen und Deregulierungen	[MUS] Die Ergebnisse des letzten Teilprojekts zeigen die Notwendigkeit, mehr staatlich geförderte Ausbildungszentren für Musiklehrer zu etablieren.
	Entwicklung von/Beiträgen zu Gesetzen/Rechtsnormen, Anpassungen bestehender Gesetze/Rechtsnormen	[MUS] Die Bedingungen für eine teilweise Weiterentwicklung von Komponenten sowie deren Verwertung werden in einem gemeinsamen Lizenzvertrag festgelegt.
	Sonstige (Other)	
Ethischer Impact	ethische Auswirkungen, z. Gleichheit, Bewusstsein, Nächstenliebe	
	Bewusstsein/Wahrnehmung/Einstellungen herstellen, z.B. Gesundheitsbewusstsein, Klimabewusstsein	[MOB] Die neuen Hambacher Forstaktivisten stützen ihren Protest auf die Ergebnisse der Klimaschutzstudie, um auf die Abholzung durch den RWE-Konzern aufmerksam zu machen.

	Verbesserung von Gerechtigkeit und Gleichstellung, einschließlich Einwanderung und Gleichstellung der Geschlechter/Teilhabe von Frauen	[LING] Wir zeigen den Einfluss der Migration auf die Entwicklung individueller und gesellschaftlicher Werte als zentrale Aspekte jugendlicher Identität auf.
	Datenschutz/Datenschutz, Open Access (aber nicht IT-Sicherheit; siehe unten)	[MOB] Die Daten werden bereits im Mobilfunknetz anonym erfasst, so dass zu keiner Zeit ein Rückschluss auf die Telefonnummer oder den Nutzer eines Mobiltelefons möglich ist.
	Sonstige (Other)	
Ökonomischer Impact	Nutzung von Forschungsergebnissen für wirtschaftliche Entwicklungen	
	Entwicklung von Geschäftsmodellen/anderen Wirtschaftsstrategien	[KI] Der Vorteil dieser Lösungen liegt darin, dass Doppel- oder Parallelentwicklungen vermieden und die Entwicklungsergebnisse und Skaleneffekte der großen Anbieter genutzt werden können, ohne proprietäre Hardware entwickeln und verkaufen zu müssen.
	Auswirkungen auf das Einkommen (außerhalb der Wissenschaft)	[MUS] Im Rahmen von Dienstleistungen geht es darum, entwickelte Einzelkomponenten in der Verarbeitung von Metadaten gegen Geldwert zu nutzen.
	Mitarbeiterzufriedenheit/ Servicequalität	[KI] Die Optimierung des Moduls führte zu einer größeren Zufriedenheit bei den Testanwendern.
	Prozessoptimierung, z.B. (Entwicklungs-)Kosten senken, Lieferketten verbessern, Markt-/Produktregulierungen etc.	[KI] Zur Ermittlung des Schadenspotenzials der Bedrohungen ist nun eine Risikomodellierung erforderlich, um qualifizierte Entscheidungen über den Einsatz des IT-Sicherheitsbudgets treffen zu können.
	Sonstige (Other)	
	Auswirkungen auf ökologische/umweltbezogene Bereiche	

FKZ 01IO2002A
Version 1.1

Ökologischer Impact	**Umwelt/Klimaschutz/Artenschutz,** inkl. Recycling/Nachwachsende Güter/Erneuerbare Energien	[MOB] Die Nutzung der Elektromobilität ist ein wichtiger Erfolgsfaktor zur Erreichung der Unabhängigkeit von fossilen Primärenergieträgern und kann auch zur dynamischen Zwischenspeicherung von regenerativ basierten Energiespitzen genutzt werden.
	Nachhaltigkeit von Produkten, Verfahren etc.	[MOB] Das R&D -Projekt ermöglicht erstmals eine praxisgerechte und wirtschaftliche Rückgewinnung und Speicherung von potentieller und kinetischer Energie, z.B. für den emissionsfreien und leisen Betrieb von Nutzfahrzeugen und Großfahrzeugen in Ballungsräumen und Innenstädten.
	Sonstige (Other)	
Technischer Impact	Technologien, die außerhalb des ursprünglichen Projekts verwendet werden	
	Entwicklung von technischen/Software-Prototypen	[MUS] Mit dem Annotationstool wurde ein ebenso effektives wie flexibles Werkzeug zur detaillierten manuellen Musikbeschreibung geschaffen.
	Modell-/Algorithmusentwicklung	[LING] Wir können einen effektiven, robusten Algorithmus zur Extraktion berufsspezifischer Informationen aus englisch- und deutschsprachigen Texten anbieten.
	Verbesserung der IT-Sicherheit	[KI] Gegenstand und Ziel des Forschungsvorhabens „IUNO" ist die Bereitstellung von Sicherheitskonzepten und -methoden, die den besonderen Anforderungen der IT-Sicherheit hinsichtlich Betrieb, Skalierbarkeit, Robustheit und Wirtschaftlichkeit in Industrie 4.0-Prozessen gerecht werden.
	Sonstige (Other)	

Akademischer Impact	Wirkung innerhalb der Wissenschaft – innerhalb oder außerhalb des eigenen Fachgebiets/der eigenen Institution	
	Auswirkungen auf das Einkommen von Forschungseinrichtungen, einschließlich Gelder für die Einstellung neuer Teammitglieder	[KI] Bei erfolgreicher Folgebewerbung könnte eine Stelle als Projektadministrator/in eingerichtet werden.
	Neue Forschungsmethoden	[LING] Wir schlagen eine neue halbautomatische Forschungsmethode zum kontrastiven Vergleich deutscher und englischer Grammatik vor.
	Neue/verbesserte Lern- und Lehrmethoden innerhalb der Wissenschaft, z. B. neue Masterprogramme, Mitarbeiterschulungen, Etablierung von Führungskompetenzen innerhalb der Wissenschaft usw.	[LING] Es lässt sich festhalten, dass sich das für die Interventionsstudie entwickelte didaktische Konzept als wirksam erweist, um den Studierenden zu helfen, ihre pragmatischen Kompetenzen, Fertigkeiten und Fähigkeiten zur Nutzung sozial konstruierter und traditioneller Handlungsmuster zu entwickeln.
	Veröffentlichung von Forschungsergebnissen, z. B. in Zeitschriften/auf Konferenzen	[MUS] Das MuBiKi-Projekt und die Ergebnisse der Evaluierung wurden in Printmedien und auf mehreren Konferenzen vorgestellt.
	Organisation von wissenschaftlichen Veranstaltungen, z. B. Workshops, Konferenzen	[MUSIC] Der Workshop für Meisterschüler des Musikinstrumentenbauhandwerks in der Manufakturregion Obervogtland war ein voller Erfolg.
	Aufbau/Verbesserung von Kooperationen/Netzwerken innerhalb der Wissenschaft	[LING] Auch im Siegener Teilprojekt des Verbundvorhabens wurde im Austausch mit dem Verbundpartner ein Modell entwickelt, das die Prozessperspektive in den Mittelpunkt rückt und wissenschaftliches Schreiben institutionell verortet.
	Sonstige (Other)	

4.3. Ad AP 3: Ergebnisse Online-Umfrage am IDS

Abbildung 4-9 Verteilung der Teilnehmenden hinsichtlich wissenschaftlicher Tätigkeit

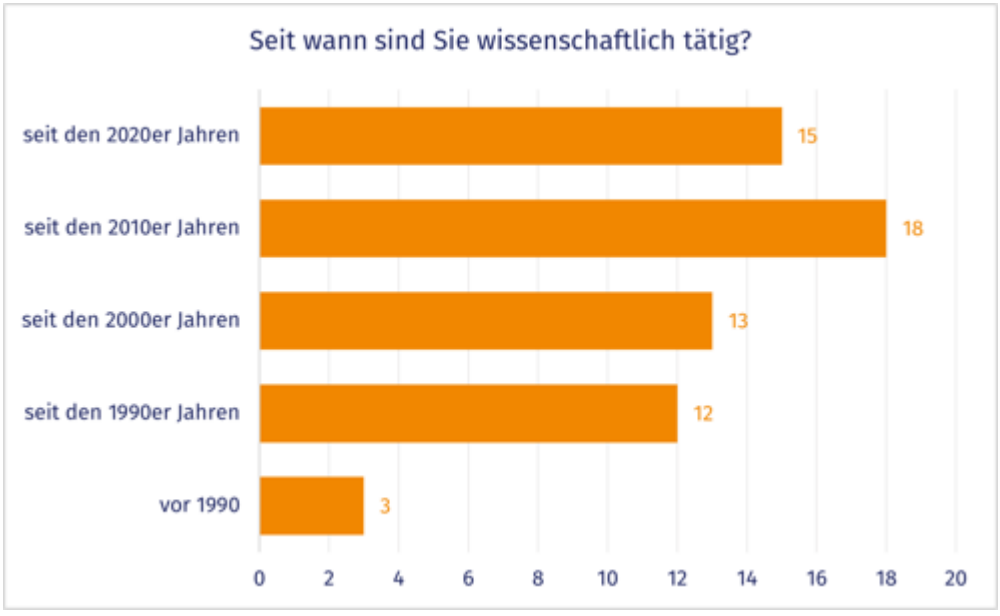

Abbildung 4-10 Dauer der wissenschaftlichen Tätigkeit

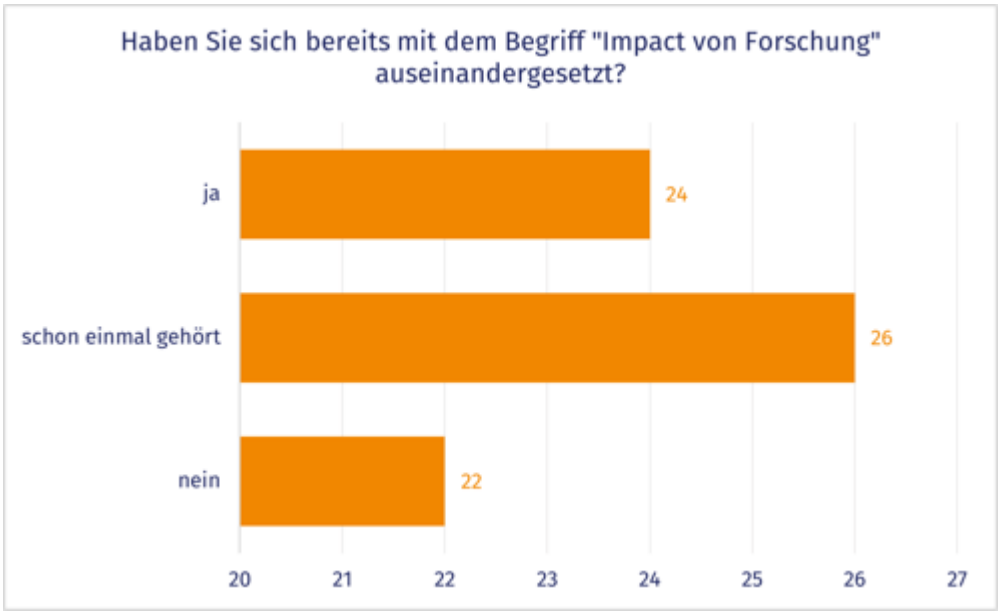

Abbildung 4-11 Vertrautheit der Teilnehmenden mit dem Begriff „Impact von Forschung“

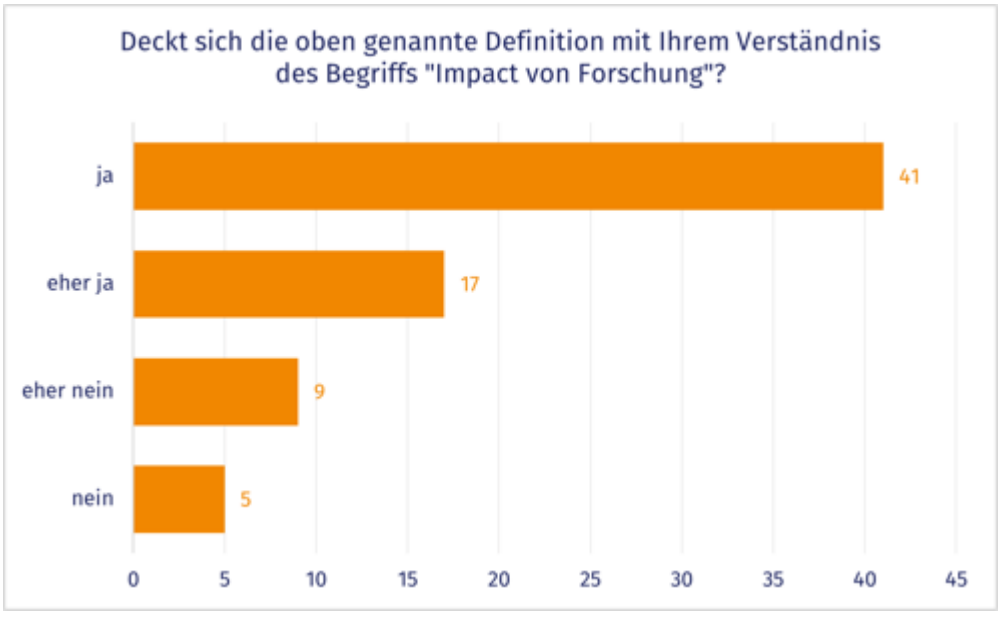

Abbildung 4-12 Verständnis der Teilnehmenden hinsichtlich „Impact von Forschung“

FKZ 01IO2002A
Version 1.1

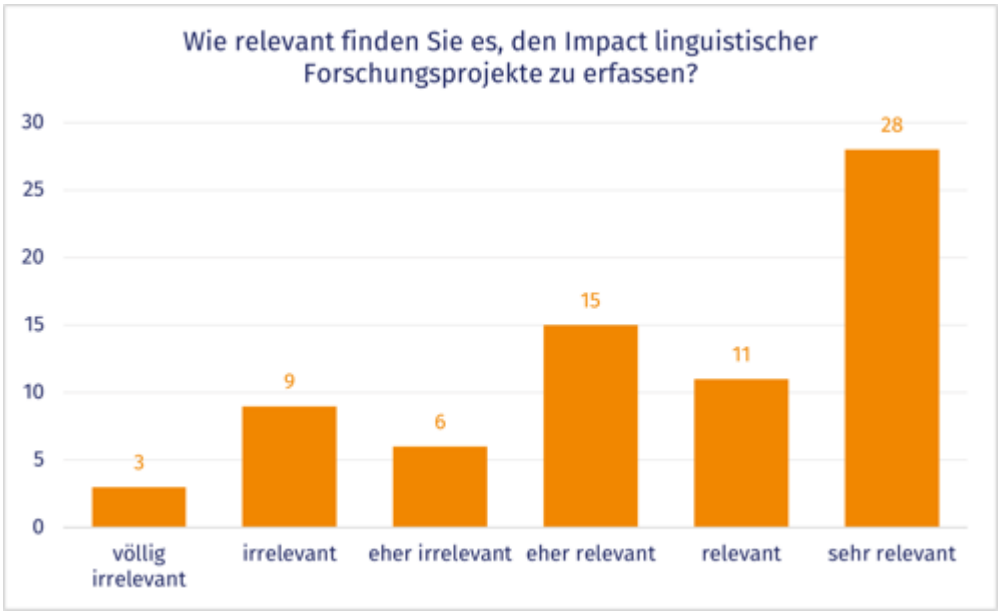

Abbildung 4-13 Wahrnehmung der Teilnehmenden hinsichtlich der Relevanz von Impacterfassung

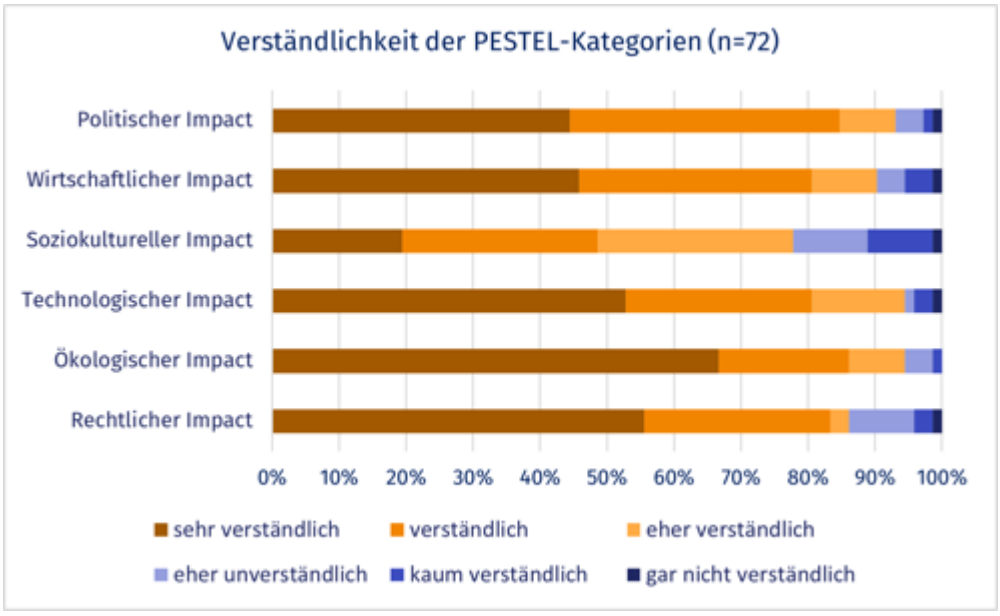

Abbildung 4-14 Verständlichkeit der PESTEL-Kategorien

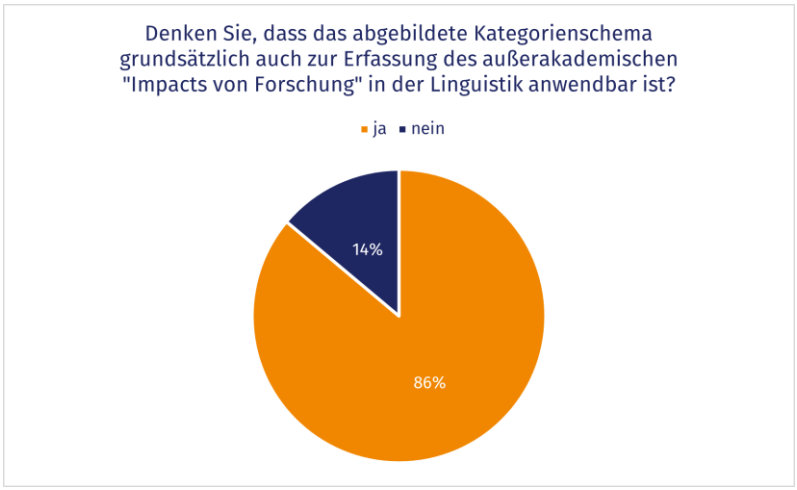

Abbildung 4-15 Eignung der PESTEL-Kategorien zur Anwendung in der Linguistik

Abbildung 4-16 Vorschläge der Teilnehmenden zur Ergänzung der Impactkategorien

FKZ 01IO2002A
Version 1.1

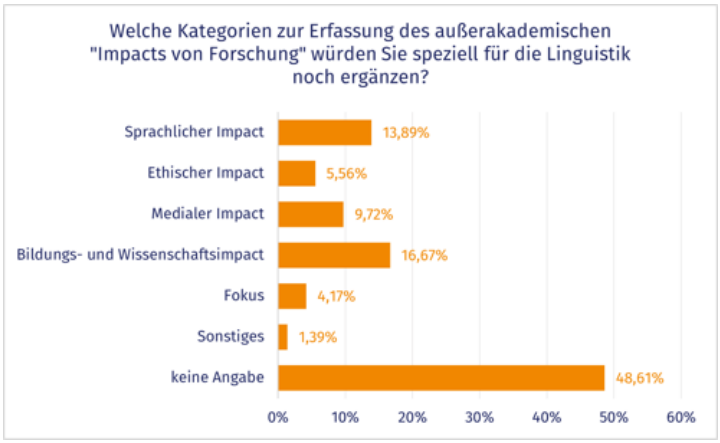

Abbildung 4-17 Verteilung der Vorschläge zur Ergänzung des Kategorienschemas

FKZ 01IO2002A
Version 1.1

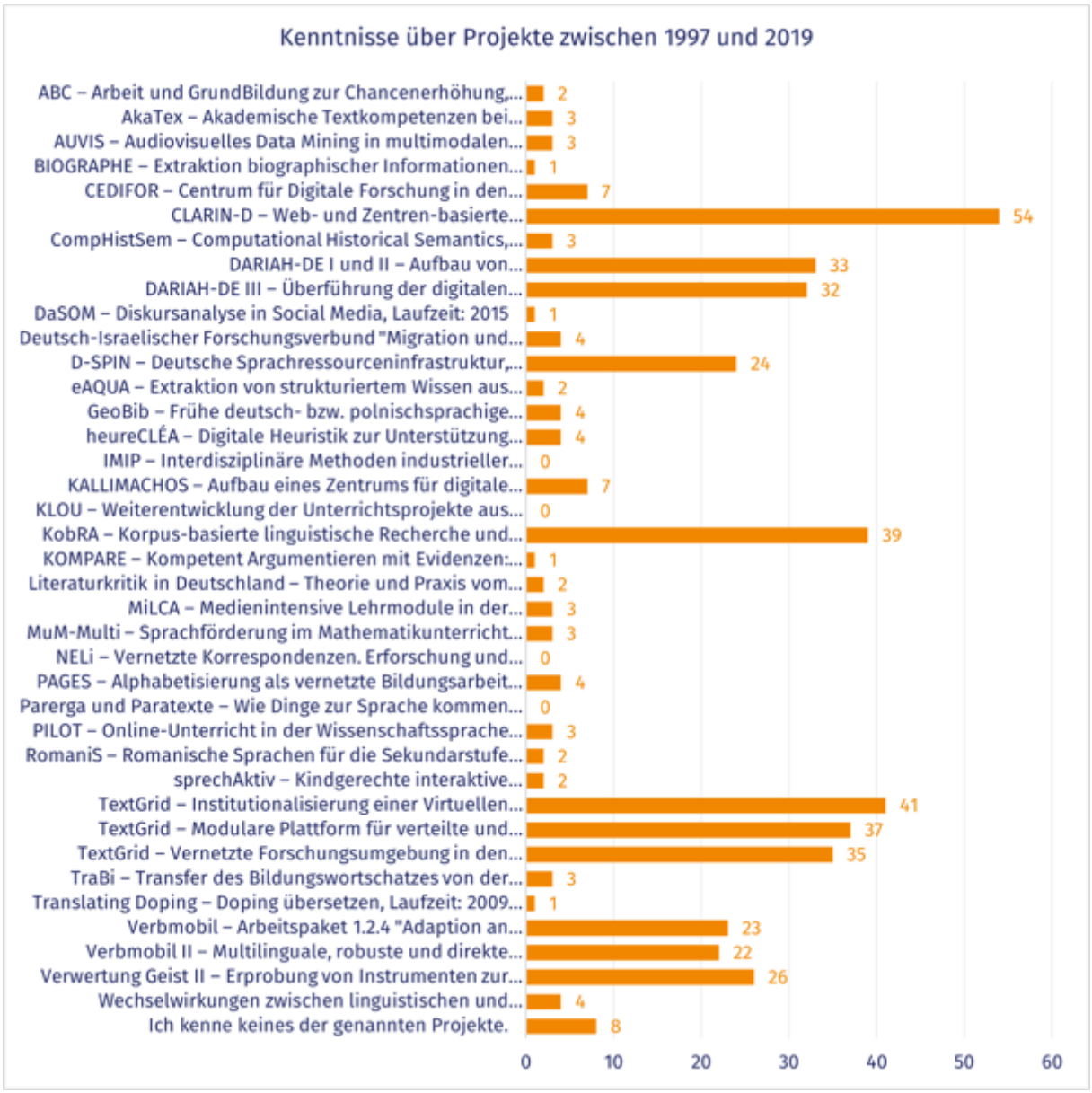

Abbildung 4-18 Kenntnisse der Teilnehmenden über einzelne Projekte aus der Linguistik aus den Jahren 1997 bis 2019

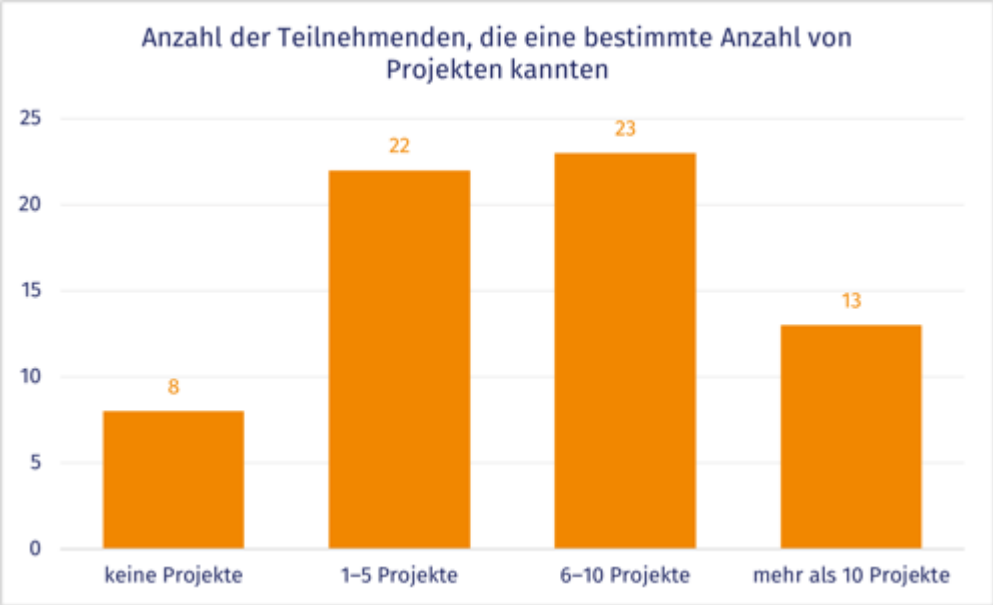

Abbildung 4-19 Anzahl der Teilnehmenden, die eine bestimmte Anzahl von Projekten kannten

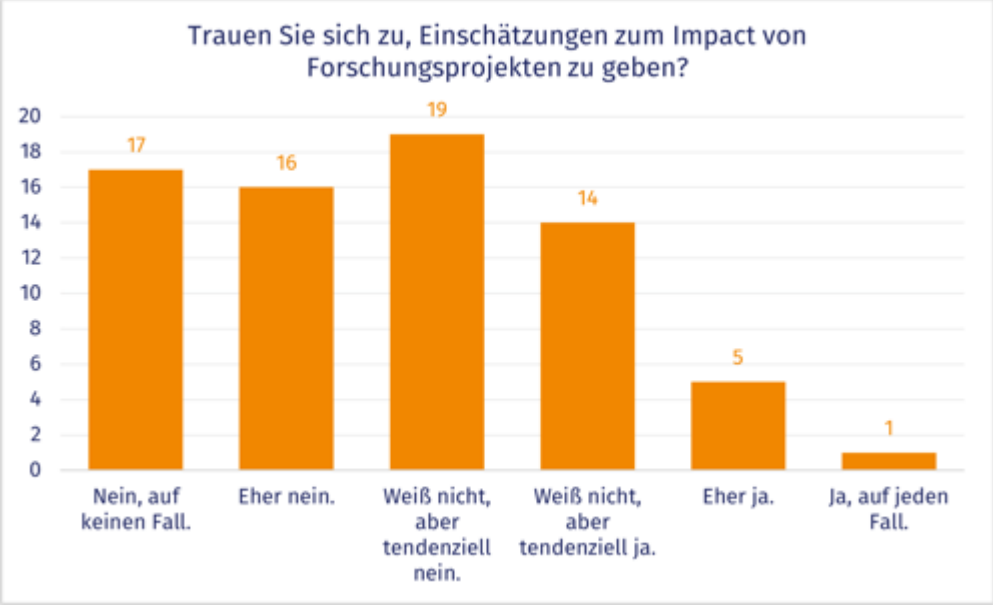

Abbildung 4-20 Zutrauen der Teilnehmenden hinsichtlich Impacteinschätzung von Forschungsprojekten

FKZ 01IO2002A
Version 1.1

Korrelationen

Die nachfolgend aufgeführten Korrelationen zwischen den verschiedenen Merkmalen bzw. Ergebnissen der Befragung sind nach *Pearson*[63] bestimmt worden.

Verständnis der PESTEL-Kategorien

Das Verständnis der PESTEL-Kategorien seitens der Teilnehmenden korreliert insgesamt signifikant positiv mit der übereinstimmenden Impactdefinition der Teilnehmenden (r beträgt pro Kategorie zwischen 0,29** und 0,35**). Dies gilt für das Verständnis sämtlicher PESTEL-Kategorien, außer für die Kategorie „Soziokultureller Impact“ (r=0,07).

Zwischen dem Verständnis der Teilnehmenden in Bezug auf die einzelnen Impactkategorien und ihrer Einschätzung hinsichtlich der Anwendbarkeit der Kategorien in der Linguistik gibt es keine signifikante Korrelation.

Anwendbarkeit der PESTEL-Kategorien

Die Meinung zur Anwendbarkeit des PESTEL-Kategorienschemas in der Linguistik ist positiv korreliert mit der Übereinstimmung der Teilnehmenden in Bezug auf die Definition des Begriffs „Impact von Forschung“ (r=0,31**).

Die Wahrnehmung der Relevanz von Impacterfassung korreliert ebenfalls positiv mit der Meinung der Teilnehmenden hinsichtlich der Anwendbarkeit des PESTEL-Kategorienschemas in der Linguistik (r=0,28**).

Gleichzeitig besteht eine positive Korrelation zwischen der Wahrnehmung der Relevanz seitens der Teilnehmenden in Bezug auf die Erfassung von Impact und der Übereinstimmung der Teilnehmenden hinsichtlich der Definition von Impact (r=0,37***).

[63] https://de.wikipedia.org/wiki/Korrelationskoeffizient

Je mehr die Impactdefinition der Teilnehmenden mit der in der Umfrage genannten Impactdefinition übereinstimmte, desto relevanter fanden sie es folglich, den außerakademischen „Impact von Forschung“ zu erfassen und umso eher waren sie der Meinung, dass das PESTEL-Kategorienschema zur Erfassung von Impact in der Linguistik angewendet werden kann.

Kenntnisse über Projekte

Die Anzahl der Projekte, die einer Person bekannt waren, ist positiv korreliert mit der Dauer der wissenschaftlichen Tätigkeit des Teilnehmenden (r=0,49***). Jedoch korreliert die Anzahl der bekannten Projekte pro Person nicht signifikant mit der Tatsache, ob diese wissenschaftlich tätig ist (r=0,18, p=0,13) oder ob diese eine leitende Funktion am IDS innehat (r=0,12, p=0,29).

Weiterhin spielt auch das Laufzeitende der Projekte keine signifikante Rolle in Bezug auf die Anzahl der bekannten Projekte pro Person (r=0,08, p=0,64).

Zutrauen zur Einschätzung von Impact

Es besteht eine leichte Korrelation zwischen der Wahrnehmung der Teilnehmenden hinsichtlich der Relevanz von Impacterfassung und der Frage, ob sie sich zutrauen würden, Einschätzungen zum Impact von Forschungsprojekten zu geben (r=0,18*).

4.4. Ad AP 3: Automatische Extraktion impactrelevanter Passagen

PDF2text

```
Pdf to Text

Please use pdf_to_text.ipynb to convert pdf files to txt files. After con-
version, you can restructure the folder as folder "texttransfer data" ->
(domain1 subfolder, domain2 subfolder, domain3 subfolder, domain4 subfolder):
each domain folder contains the text files of reports (one text file per
report). Only in this way can you run the second notebook (do not forget to
modify the dir you store the "texttransfer data" folder).
```

```
import os

import re

import pandas as pd

path = 'C:/Users/kanya/Desktop/Documents/texttransfer data'

subfolder = os.listdir(path)
```

```
all_docs = []

domains = []

projects = []

files = []

for i in subfolder:
```

```
    for j in os.listdir(path+'/'+ i):
        with open (path+'/'+ i + '/' + j, "r", encoding="utf-8") as myfile:
            data=myfile.readlines()
            data = ' '.join(data)
        all_docs += [data]
        domains += [i]
        ind = [m.start() for m in re.finditer('_', j)]
        files += [j]
        projects += [j[ind[-2]+1:ind[-1]]]



data = pd.DataFrame({'Domain': domains,
                     'Projects': projects,
                     'file':files,
                     'text': all_docs})



data['text1'] = data['text'].str.replace('-\n ', '').str.replace('-\n', '')

data['text1'] = data['text1'].apply(lambda x: re.sub('\s+pagebreak', ' PAGE-
BREAK wozhizhi', x))

data['text1'] = data['text1'].str.replace('\n \n', ' wozhidaode').str.re-
place('\n  \n', ' wozhidaode').str.replace('\n \d+', ' wozhidaode').str.re-
place('\n\d+',       '       wozhidaode').str.replace('\n',       '').str.re-
place('wozhidaode', '.\n \n')
```

```
data['text2'] = data['text1'].str.replace('. \n ', '.\n ').str.replace('.\n
', 'wozhizhi').str.split('wozhizhi')
```

```
C:\Users\kanya\AppData\Local\Temp/ipykernel_277340/3152228529.py:3: Future-
Warning: The default value of regex will change from True to False in a
future version.

  data['text1'] = data['text1'].str.replace('\n \n', ' wozhidaode').str.re-
place('\n  \n', ' wozhidaode').str.replace('\n \d+', ' wozhidaode').str.re-
place('\n\d+',       '       wozhidaode').str.replace('\n',       '').str.re-
place('wozhidaode', '.\n \n')

C:\Users\kanya\AppData\Local\Temp/ipykernel_277340/3152228529.py:4: Future-
Warning: The default value of regex will change from True to False in a
future version.

  data['text2'] = data['text1'].str.replace('. \n ', '.\n ').str.replace('.\n
', 'wozhizhi').str.split('wozhizhi')
```

```
for i in range(len(data)):

    if i == 0:

        re_data = pd.DataFrame({'text': data['text2'][i]})

        re_data['Domain'] = data['Domain'][i]

        re_data['Projects'] = data['Projects'][i]

        re_data['file'] = data['file'][i]

    else:

        re_data1 = pd.DataFrame({'text': data['text2'][i]})
```

```
        re_data1['Domain'] = data['Domain'][i]
        re_data1['Projects'] = data['Projects'][i]
        re_data1['file'] = data['file'][i]
        re_data = pd.concat([re_data, re_data1])



re_data.index = range(len(re_data))
re_data['text'] = re_data['text'].str.replace('\s+', ' ', regex=True)



re_data['len'] = re_data['text'].apply(lambda x: len(x))
re_data = re_data[re_data['len'] >6]
re_data.index = range(len(re_data))



re_data['text'] = re_data['text'].apply(lambda x: x[1:] if x[0] == ' ' else
x)
re_data['text'] = re_data['text'].apply(lambda x: x[:-1] if x[-1] == ' ' else
x)



re_data['len'] = re_data['text'].apply(lambda x: len(x))
re_data = re_data[re_data['len'] >5]
```


FKZ 01IO2002A
Version 1.1

```
re_data.index = range(len(re_data))



re_data['text4model'] = re_data['text']







tokens = re_data['text4model'].str.split(' ')
for i in range(len(data)):
    for j in range(len(tokens[i])):
        if tokens[i][j].isupper():
            tokens[i][j] = tokens[i][j].title()
re_data['text4model'] = tokens
re_data['text4model'] = re_data['text4model'].apply(lambda x: ' '.join(x))



re_data['text4model']   =   re_data['text4model'].str.replace('\.\.\.',   '
').str.replace('\s+', ' ', regex=True)



  re_data['text4model']   =   re_data['text4model'].str.replace('\.\.\.',   '
').str.replace('\s+', ' ', regex=True)
```


FKZ 01IO2002A
Version 1.1

```
re_data['text4model'] = re_data['text4model'].apply(lambda x: ' '.join([to-
ken for token in x.split() if len(token) >= 2]))



for text in re_data['text4model']:
    doc = nlp(text)
    result = " ".join([ent.lemma_ for ent in doc])
    lemma.append(result)



re_data['text4model'] = lemma



from nltk.corpus import stopwords
stop = set(stopwords.words())
re_data['text4model'] = re_data['text4model'].apply(lambda x: ' '.join([to-
ken for token in x.split() if token not in stop]))
re_data['text4model'] = re_data['text4model'].apply(lambda x: ' '.join([to-
ken for token in x.split() if len(token) >= 3]))
import string
for i in range(len(data)):
    re_data['text4model'][i] = re_data['text4model'][i].translate(str.maket-
rans('', '', string.punctuation))
```

```
re_data['text4model'] = re_data['text4model'].str.replace('\s+', ' ', re-
gex=True)



re_data['id'] = re_data.index

re_data.to_csv('C:/Users/kanya/Desktop/ttparagraph_addmob.txt.gz', encoding
= 'utf8', index = False)




sample_list = ['KI_737_AutoPNP_856753408.txt',
               'KI_517_Saki_863856837.txt',
               'KI_578_SmartDataWeb_1686069537.txt',
               'KI_320_IUNO_1670745120.txt',
               'KI_690_UHCI_869843966.txt',
               'Ling_126_AkaTex_873190904.txt',
               'Ling_147_BIOGRAPHE_776146335.txt',
               'Ling_207_Literaturkritik_487390423.txt',
               'Ling_161_TextGridVernetzteForschungsumge-
bung_768470994.txt',
               'Ling_196_DeutschIsraelischer_667762345.txt',
               'MuWi_014_GlobalMusic2one_719934826.txt',
               'MuWi_120_InnoRegio Musicon Valley_487476190.txt',
```

```
                'MuWi_051_Kompetenznetzwerk QM & LE_1015232051.txt',
                'MuWi_049_MuBiKi_1024722414.txt',
                'MuWi_034_DysTract_1693406810.txt']


re_data[re_data['file'].isin(sample_list)].to_excel('C:/Users/kanya/Desk-
top/impact_paragraph.xlsx', encoding = 'utf8', index = False)


pd.read_excel('C:/Users/kanya/Desktop/impact_paragraph.xlsx')
```


FKZ 01IO2002A
Version 1.1

Data Cleaning

data clean and restructure

Data cleaning and restructuring.ipynb is a notebook for data cleaning and restructure. It will read all text files and split them into paragraphs. It will generate two dataset: ttparagraph_addmob.txt.gz is the corpus while impact_paragraph.xlsx is the pilot study data (for annotation and model training).

```
import os

import re

import pandas as pd

path = 'C:/Users/kanya/Desktop/Documents/texttransfer data'

subfolder = os.listdir(path)



all_docs = []

domains = []

projects = []

files = []

for i in subfolder:

    for j in os.listdir(path+'/'+ i):

        with open (path+'/'+ i + '/' + j, "r", encoding="utf-8") as myfile:
```


FKZ 01IO2002A
Version 1.1

```
            data=myfile.readlines()
            data = ' '.join(data)
        all_docs += [data]
        domains += [i]
        ind = [m.start() for m in re.finditer('_', j)]
        files += [j]
        projects += [j[ind[-2]+1:ind[-1]]]



data = pd.DataFrame({'Domain': domains,
                     'Projects': projects,
                     'file':files,
                     'text': all_docs})



data['text1'] = data['text'].str.replace('-\n ', '').str.replace('-\n', '')
data['text1'] = data['text1'].apply(lambda x: re.sub('\s+pagebreak', ' PAGE-
BREAK wozhizhi', x))
data['text1'] = data['text1'].str.replace('\n \n', ' wozhidaode').str.re-
place('\n  \n', ' wozhidaode').str.replace('\n \d+', ' wozhidaode').str.re-
place('\n\d+',      '      wozhidaode').str.replace('\n',      '').str.re-
place('wozhidaode', '.\n \n')
data['text2'] = data['text1'].str.replace('. \n ', '.\n ').str.replace('.\n
', 'wozhizhi').str.split('wozhizhi')
```

```
C:\Users\kanya\AppData\Local\Temp/ipykernel_277340/3152228529.py:3: Future-
Warning: The default value of regex will change from True to False in a
future version.

  data['text1'] = data['text1'].str.replace('\n \n', ' wozhidaode').str.re-
place('\n  \n', ' wozhidaode').str.replace('\n \d+', ' wozhidaode').str.re-
place('\n\d+',       '      wozhidaode').str.replace('\n',       '').str.re-
place('wozhidaode', '.\n \n')

C:\Users\kanya\AppData\Local\Temp/ipykernel_277340/3152228529.py:4: Future-
Warning: The default value of regex will change from True to False in a
future version.

  data['text2'] = data['text1'].str.replace('. \n ', '.\n ').str.replace('.\n
', 'wozhizhi').str.split('wozhizhi')


for i in range(len(data)):

    if i == 0:

        re_data = pd.DataFrame({'text': data['text2'][i]})

        re_data['Domain'] = data['Domain'][i]

        re_data['Projects'] = data['Projects'][i]

        re_data['file'] = data['file'][i]

    else:

        re_data1 = pd.DataFrame({'text': data['text2'][i]})

        re_data1['Domain'] = data['Domain'][i]
```


FKZ 01IO2002A
Version 1.1

```
        re_data1['Projects'] = data['Projects'][i]
        re_data1['file'] = data['file'][i]
        re_data = pd.concat([re_data, re_data1])


re_data.index = range(len(re_data))
re_data['text'] = re_data['text'].str.replace('\s+', ' ', regex=True)


re_data['len'] = re_data['text'].apply(lambda x: len(x))
re_data = re_data[re_data['len'] >6]
re_data.index = range(len(re_data))


re_data['text'] = re_data['text'].apply(lambda x: x[1:] if x[0] == ' ' else
x)
re_data['text'] = re_data['text'].apply(lambda x: x[:-1] if x[-1] == ' ' else
x)


re_data['len'] = re_data['text'].apply(lambda x: len(x))
re_data = re_data[re_data['len'] >5]
re_data.index = range(len(re_data))
```


FKZ 01IO2002A
Version 1.1

```
re_data['text4model'] = re_data['text']
```

```
tokens = re_data['text4model'].str.split(' ')
for i in range(len(data)):
    for j in range(len(tokens[i])):
        if tokens[i][j].isupper():
            tokens[i][j] = tokens[i][j].title()
re_data['text4model'] = tokens
re_data['text4model'] = re_data['text4model'].apply(lambda x: ' '.join(x))
```

```
re_data['text4model']   =   re_data['text4model'].str.replace('\.\.\.',   '
').str.replace('\s+', ' ', regex=True)
```

```
C:\Users\kanya\AppData\Local\Temp/ipykernel_269280/3536642181.py:1: Future-
Warning: The default value of regex will change from True to False in a
future version.
  re_data['text4model']   =   re_data['text4model'].str.replace('\.\.\.',   '
').str.replace('\s+', ' ', regex=True)
```


FKZ 01IO2002A
Version 1.1

```
re_data['text4model'] = re_data['text4model'].apply(lambda x: ' '.join([to-
ken for token in x.split() if len(token) >= 2]))



for text in re_data['text4model']:
    doc = nlp(text)
    result = " ".join([ent.lemma_ for ent in doc])
    lemma.append(result)



re_data['text4model'] = lemma



from nltk.corpus import stopwords
stop = set(stopwords.words())
re_data['text4model'] = re_data['text4model'].apply(lambda x: ' '.join([to-
ken for token in x.split() if token not in stop]))
re_data['text4model'] = re_data['text4model'].apply(lambda x: ' '.join([to-
ken for token in x.split() if len(token) >= 3]))
import string
for i in range(len(data)):
    re_data['text4model'][i] = re_data['text4model'][i].translate(str.maket-
rans('', '', string.punctuation))
```

```
re_data['text4model'] = re_data['text4model'].str.replace('\s+', ' ', re-
gex=True)



re_data['id'] = re_data.index

re_data.to_csv('C:/Users/kanya/Desktop/ttparagraph_addmob.txt.gz', encoding
= 'utf8', index = False)



sample_list = ['KI_737_AutoPNP_856753408.txt',

               'KI_517_Saki_863856837.txt',

               'KI_578_SmartDataWeb_1686069537.txt',

               'KI_320_IUNO_1670745120.txt',

               'KI_690_UHCI_869843966.txt',

               'Ling_126_AkaTex_873190904.txt',

               'Ling_147_BIOGRAPHE_776146335.txt',

               'Ling_207_Literaturkritik_487390423.txt',

               'Ling_161_TextGridVernetzteForschungsumge-
bung_768470994.txt',

               'Ling_196_DeutschIsraelischer_667762345.txt',

               'MuWi_014_GlobalMusic2one_719934826.txt',

               'MuWi_120_InnoRegio Musicon Valley_487476190.txt',

               'MuWi_051_Kompetenznetzwerk QM & LE_1015232051.txt',

               'MuWi_049_MuBiKi_1024722414.txt',
```

```
'MuWi_034_DysTract_1693406810.txt']
```

```
re_data[re_data['file'].isin(sample_list)].to_excel('C:/Users/kanya/Desk-
top/impact_paragraph.xlsx', encoding = 'utf8', index = False)
```

```
pd.read_excel('C:/Users/kanya/Desktop/impact_paragraph.xlsx')
```

Paragraph extraction with random forest

```
paragraph extraction with random forest

paragraph_extraction.ipynb is a "long" notebook for paragraph extraction,
including rule-based extraction and random forest model training and predic-
tion, as well as datasets merging (TT-I + TT-II).


import pandas as pd

import math



##Random forest model training


## past annotated data in the first phase of TT project

past_data = pd.read_csv('./annotated_data.big_set.corrected.txt', sep=';' ,
encoding='latin-1',

                        header = None,

                        names = ['doc', 'col1', 'col2', 'text', 'label1',
'label2', 'label3'])



## annotated data by the German team

data = pd.read_excel('./evaluation_20220927.ods', engine = 'odf')
```


FKZ 01IO2002A
Version 1.1

```
## identify table of content

data['inhalt1']      =      data['text'].str.contains('Inhaltsverzeichnis',
case=True).apply(lambda x: 1 if x else 0)

data['inhalt2']    =    data['text'].str.contains('Inhalt',    case=True).ap-
ply(lambda x: 1 if x else 0)

inhalt1_sum = data.groupby('file').sum('inhalt1')

inhalt1_ind = inhalt1_sum[inhalt1_sum['inhalt1'] > 0].index.tolist()

inhalt2_ind = set(inhalt1_sum.index) - set(inhalt1_ind)


data['exclude_inhalt'] = False



list_exclude = []

for i in inhalt1_ind:

    data_part = data[data['file'] == i]

    ind_start = data_part.index.tolist()[0]

    data_part = data_part[data_part['inhalt1'] == True]

    if len(data_part) > 0:

        ind_end = data_part[data_part['inhalt1'] == True].index[0]

        range_inhalt = list(range(ind_start, ind_end+1))

        if len(range_inhalt) <= 30: # change
```

```
            list_exclude = list_exclude + range_inhalt


for i in inhalt2_ind:
    data_part = data[data['file'] == i]
    ind_start = data_part.index.tolist()[0]
    data_part = data_part[data_part['inhalt2'] == True]
    if len(data_part) > 0:
        ind_end = data_part[data_part['inhalt2'] == True].index[0]
        range_inhalt = list(range(ind_start, ind_end+1))
        if len(range_inhalt) <= 30: # change
            list_exclude = list_exclude + range_inhalt



data['exclude_inhalt'][list_exclude] = True



/tmp/ipykernel_679427/2894172796.py:33: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame


See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data['exclude_inhalt'][list_exclude] = True
```


FKZ 01IO2002A
Version 1.1

```
# identify bibliography


data['inhalt1'] = data['text'].str.contains('Bibliographie|Bibliografie|Lit-
eraturverzeichnis', case=True).apply(lambda x: 1 if x else 0)

data['inhalt2']  =  data['text'].str.contains('Literatur',  case=True).ap-
ply(lambda x: 1 if x else 0)

inhalt1_sum = data.groupby('file').sum('inhalt1')

inhalt1_ind = inhalt1_sum[inhalt1_sum['inhalt1'] > 0].index.tolist()

inhalt2_ind = set(inhalt1_sum.index) - set(inhalt1_ind)

data['exclude_inhalt1'] = False



list_exclude = []

for i in inhalt1_ind:

    data_part = data[data['file'] == i]

    ind_start = data_part.index.tolist()[-1]

    data_part = data_part[data_part['inhalt1'] == True]

    if len(data_part) > 0:

        ind_end = data_part.index[-1]

        range_inhalt = list(range(ind_end, ind_start+1))

        if len(range_inhalt) <= 75: # change

            list_exclude = list_exclude + range_inhalt
```

```
for i in inhalt2_ind:
    data_part = data[data['file'] == i]
    ind_start = data_part.index.tolist()[-1]
    data_part = data_part[data_part['inhalt2'] == True]
    if len(data_part) > 0:
        ind_end = data_part.index[-1]
        range_inhalt = list(range(ind_end, ind_start+1))
        if len(range_inhalt) <= 75: # change
            list_exclude = list_exclude + range_inhalt



data['exclude_inhalt1'][list_exclude] = True
```

```
/tmp/ipykernel_679427/2453373711.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data['exclude_inhalt1'][list_exclude] = True
```

```
# calculate the number of token in each paragraph and remove those with too less tokens
```

```
data['text'] = data['text'].apply(lambda x: str(x))
```

```
data['num_token'] = data['text'].str.replace('.', '').str.replace(r"\s+", '
').apply(lambda x: len(x.split(' ')))
```

```
/tmp/ipykernel_679427/2130238111.py:1: FutureWarning: The default value of
regex will change from True to False in a future version. In addition, single
character regular expressions will *not* be treated as literal strings when
regex=True.
  data['num_token'] = data['text'].str.replace('.', '').str.replace(r"\s+",
' ').apply(lambda x: len(x.split(' ')))
/tmp/ipykernel_679427/2130238111.py:1: FutureWarning: The default value of
regex will change from True to False in a future version.
  data['num_token'] = data['text'].str.replace('.', '').str.replace(r"\s+",
' ').apply(lambda x: len(x.split(' ')))
```

```
data['exclude_num_token'] = (data['num_token'] < 25) # change
```

```
# remove all irrelevant paragraphs identifed by rule based methods above
```

```
data = data[data['exclude_inhalt'] == False][data['exclude_inhalt1'] == False][data['exclude_num_token'] == False]
```

```
/tmp/ipykernel_679427/695616058.py:1: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  data = data[data['exclude_inhalt'] == False][data['exclude_inhalt1'] == False][data['exclude_num_token'] == False]
```

```
# Since the TT-I annotated data is at sentence level, merge them into paragraph level based on the average length of paragraphs in our current data
text = []
doc = []
for i in past_data['doc'].unique():
    df = past_data[past_data['doc'] == i]
    length = math.ceil(len(df)/5)
    for h in range(length):
        text.append(df[h*5:(h+1)*5]['text'].str.cat(sep = ' '))
        doc.append(i)
```

```
# Merge the past and current datasets
```

```
df1 = data[['text', 'label_update_with_scientific_impact', 'test']]
```


FKZ 01IO2002A
Version 1.1

```
df2 = pd.DataFrame({'text': text, 'label_update_with_scientific_impact':
[1]*len(text),
                    'test': ['no']*len(text)})



data = pd.concat([df1, df2])

data.index = range(len(data))



# keywords

cate1 = ['auswirken', 'Auswirkung', 'beeinflussen', 'beinflußen',

 'Effekt', 'effektiv', 'Einfluss', 'Einfluß', 'Fortschritt', 'Impact', 'nach-
haltig', 'nutzbar', 'Nutzbarmachung',

 'Potential', 'Potenzial', 'umsetzen', 'Umsetzung', 'verändern', 'Verän-
derung', 'verbessern',

 'Verbesserung', 'Verwertung', 'Verwertungsmöglichkeiten', 'wirksam', 'Wirk-
samkeit', 'Wirkung']

cate2 = ['beachtlich', 'Beitrag', 'beitragen', 'direkt', 'Einflussnahme',
'Einflußnahme', 'Einflußmöglichkeit',

         'Einflussmöglichkeit', 'Einsatzmöglichkeiten', 'hochrelevant',

         'Innovation', 'innovativ', 'realisierbar', 'realisieren', 'Reali-
sierung', 'Ziel', 'zielführend']

cate3= 'abschätzbar, abschätzen, anwenden, Anwendung, Anwendungsfall, Anwen-
dungsframework, Anwendungsszenario, Attraktivität, effizient, Entwicklung,
Erfolg, Erfolgsaussichten, Ergebnisse, ermöglichen, erreichen, erzielen,
```


FKZ 01IO2002A
Version 1.1

```
Feedback, Frontend, Gewinn, gewinnen, gewinnorientiert, Hauptanwen-
dungsfälle, indirekt, Infrastruktur, infrastrukturell, langfristig, lösen,
Lösung, maßgeblich, messbar, meßbar, negativ, neu, nutzen, positiv, produk-
tiv, Projektziele, reagieren, Reaktion, real-world, spürbar, strukturell,
Überwindung, unmittelbar, Use_Case, Weiterentwicklung, Wertschöpfung, Wett-
bewerb, Wettbewerbsanalyse, Zukunft, zukünftig, Zweck'.split(', ')



# compute the frequencies of each keyword in texts as well as the aggregrated
number

data['aaa'] = data['text'].apply(lambda x: len([word for word in x.split()
if word in set(cate1)]))

data['bbb'] = data['text'].apply(lambda x: len([word for word in x.split()
if word in set(cate2)]))

data['ccc'] = data['text'].apply(lambda x: len([word for word in x.split()
if word in set(cate3)]))



data['total'] = data['aaa'] + data['bbb'] + data['ccc']

for i in set(cate1+cate2+cate3):

    data[str(i)] = data['text'].apply(lambda x: x.count(i))



/tmp/ipykernel_679427/2400321340.py:3: PerformanceWarning: DataFrame is
highly fragmented. This is usually the result of calling `frame.insert` many
times, which has poor performance. Consider joining all columns at once
```

```
using pd.concat(axis=1) instead. To get a de-fragmented frame, use `newframe
= frame.copy()`

  data[str(i)] = data['text'].apply(lambda x: x.count(i))

/tmp/ipykernel_679427/2400321340.py:3: PerformanceWarning: DataFrame is
highly fragmented.  This is usually the result of calling `frame.insert` many
times, which has poor performance.  Consider joining all columns at once
using pd.concat(axis=1) instead. To get a de-fragmented frame, use `newframe
= frame.copy()`

  data[str(i)] = data['text'].apply(lambda x: x.count(i))


# training - validation/testing set split

train = data[data['test'] == 'no']

test = data[data['test'] == 'yes']



len(train)



2054


len(test)
```

```
444


#shuffle data

train = train.sample(frac = 1, random_state = 10)



# best model parameters and result

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report

iid = data.columns[3:]

clf = RandomForestClassifier(n_estimators = 1000, class_weight = 'balanced',
min_samples_leaf   =   6,   max_depth=6).fit(train[iid],   train['label_up-
date_with_scientific_impact'])

final = clf.predict(test[iid])

print(classification_report(test['label_update_with_scientific_impact'],
final))
```

	precision	recall	f1-score	support
0	0.85	0.67	0.75	331
1	0.41	0.66	0.51	113

FKZ 01IO2002A
Version 1.1

```
    accuracy                           0.67       444
   macro avg       0.63      0.67      0.63       444
weighted avg       0.74      0.67      0.69       444
```

```
Prediction
```

```
data = pd.read_csv('./ttparagraph_addmob.txt.gz')
```

```
Using rule-based method to identify all irrelevant paragraphs
```

```
data['inhalt1']      =      data['text'].str.contains('Inhaltsverzeichnis',
case=True).apply(lambda x: 1 if x else 0)

data['inhalt2']    =    data['text'].str.contains('Inhalt',    case=True).ap-
ply(lambda x: 1 if x else 0)

inhalt1_sum = data.groupby('file').sum('inhalt1')

inhalt1_ind = inhalt1_sum[inhalt1_sum['inhalt1'] > 0].index.tolist()

inhalt2_ind = set(inhalt1_sum.index) - set(inhalt1_ind)
```

```
data['exclude_inhalt'] = False
```

```
list_exclude = []
```

```
for i in inhalt1_ind:
    data_part = data[data['file'] == i]
    ind_start = data_part.index.tolist()[0]
    data_part = data_part[data_part['inhalt1'] == True]
    if len(data_part) > 0:
        ind_end = data_part[data_part['inhalt1'] == True].index[0]
        range_inhalt = list(range(ind_start, ind_end+1))
        if len(range_inhalt) <= 30: # change
            list_exclude = list_exclude + range_inhalt


for i in inhalt2_ind:
    data_part = data[data['file'] == i]
    ind_start = data_part.index.tolist()[0]
    data_part = data_part[data_part['inhalt2'] == True]
    if len(data_part) > 0:
        ind_end = data_part[data_part['inhalt2'] == True].index[0]
        range_inhalt = list(range(ind_start, ind_end+1))
        if len(range_inhalt) <= 30: # change
            list_exclude = list_exclude + range_inhalt



data['exclude_inhalt'][list_exclude] = True
/tmp/ipykernel_656949/2894172796.py:33: SettingWithCopyWarning:
```

```
A value is trying to be set on a copy of a slice from a DataFrame


See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data['exclude_inhalt'][list_exclude] = True


data['inhalt1'] = data['text'].str.contains('Bibliographie|Bibliografie|Lit-
eraturverzeichnis', case=True).apply(lambda x: 1 if x else 0)

data['inhalt2']  =  data['text'].str.contains('Literatur',  case=True).ap-
ply(lambda x: 1 if x else 0)

inhalt1_sum = data.groupby('file').sum('inhalt1')

inhalt1_ind = inhalt1_sum[inhalt1_sum['inhalt1'] > 0].index.tolist()

inhalt2_ind = set(inhalt1_sum.index) - set(inhalt1_ind)

data['exclude_inhalt1'] = False



list_exclude = []

for i in inhalt1_ind:

    data_part = data[data['file'] == i]

    ind_start = data_part.index.tolist()[-1]

    data_part = data_part[data_part['inhalt1'] == True]

    if len(data_part) > 0:

        ind_end = data_part.index[-1]

        range_inhalt = list(range(ind_end, ind_start+1))
```


FKZ 01IO2002A
Version 1.1

```
        if len(range_inhalt) <= 75: # change
            list_exclude = list_exclude + range_inhalt

for i in inhalt2_ind:
    data_part = data[data['file'] == i]
    ind_start = data_part.index.tolist()[-1]
    data_part = data_part[data_part['inhalt2'] == True]
    if len(data_part) > 0:
        ind_end = data_part.index[-1]
        range_inhalt = list(range(ind_end, ind_start+1))
        if len(range_inhalt) <= 75: # change
            list_exclude = list_exclude + range_inhalt


data['exclude_inhalt1'][list_exclude] = True


/tmp/ipykernel_656949/2453373711.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data['exclude_inhalt1'][list_exclude] = True
```

```
data['text'] = data['text'].apply(lambda x: str(x))
```

```
data['num_token'] = data['text'].str.replace('.', '').str.replace(r"\s+", '
').apply(lambda x: len(x.split(' ')))
```

```
/tmp/ipykernel_656949/2130238111.py:1: FutureWarning: The default value of
regex will change from True to False in a future version. In addition, single
character regular expressions will *not* be treated as literal strings when
regex=True.
  data['num_token'] = data['text'].str.replace('.', '').str.replace(r"\s+",
' ').apply(lambda x: len(x.split(' ')))
/tmp/ipykernel_656949/2130238111.py:1: FutureWarning: The default value of
regex will change from True to False in a future version.
  data['num_token'] = data['text'].str.replace('.', '').str.replace(r"\s+",
' ').apply(lambda x: len(x.split(' ')))
```

```
data['exclude_num_token'] = (data['num_token'] < 25) # change
```

```
Using the random to predict and concatenate rule-based and model prediction
result
```

```
final1 = clf.predict(data[data.columns[12:]])
```

```
data['relevant'] = final1
```

```
aa = data[['text','Domain','Projects', 'file', 'exclude_inhalt', 'exclude_inhalt1', 'exclude_num_token', 'relevant']]
```

```
ind = aa[aa['exclude_inhalt'] == False][aa['exclude_inhalt1'] == False][aa['exclude_num_token'] == False][aa['relevant']==1].index
```

```
/tmp/ipykernel_656949/1710346095.py:1: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  ind = aa[aa['exclude_inhalt'] == False][aa['exclude_inhalt1'] == False][aa['exclude_num_token'] == False][aa['relevant']==1].index
```

```
aa['relevant'] = 0
```

```
/tmp/ipykernel_656949/2946407759.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

```
Try using .loc[row_indexer,col_indexer] = value instead


See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  aa['relevant'] = 0


relevant = []
for i in range(len(aa)):
    if i in ind:
        relevant.append(1)
    else:
        relevant.append(0)



aa['relevant'] = relevant



/tmp/ipykernel_656949/1965354603.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead


See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
  aa['relevant'] = relevant


# save results
aa[['text','Domain','Projects',    'file',    'relevant']].to_csv('model_re-
sult_extraction.csv')
```

4.5. Ad AP 3: Identifikation externer Impactreferenzen - Python Scripts

```
# Script, das aus den Projektbeschreibungen alle Inhaltswörter extra-
hiert (anwendbar zur Generierung komplexer Suchausdrücke, AP "Identi-
fikation externer Impactreferenzen")

# Autorin: Maria Becker


import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word_tokenize

import sys

from stop_words import get_stop_words


with open(sys.argv[1]) as in_file:

    for line in in_file:

        stop_words = get_stop_words('german')

        word_tokens = word_tokenize(line)

        filtered_sentence = [w for w in word_tokens if not w.lower()
in stop_words]

        filtered_sentence = []

        for w in word_tokens:

            if w not in stop_words:

                filtered_sentence.append(w)

        print(filtered_sentence)


# Script, das die Daten aus den Projektberichten als passenden Input
für das ToolGermaNER (https://github.com/tudarmstadt-lt/GermaNER)
konvertiert, AP "Identifikation externer Impactreferenzen")

# Autorin: Maria Becker
```

```
import fileinput
import re
import sys
with open(sys.argv[1]) as test_text:
    data = test_text.read()
    filtered2 = re.sub("\n", "", data)
    filtered3 = re.sub("\s\s"," ", filtered2)
    filtered4 = re.sub("\s\s"," ", filtered3)
    filtered5 = re.sub("\s\s"," ", filtered4)
    filtered6 = re.sub("\s\s"," ", filtered5)
    filtered7 = re.sub("\s\s"," ", filtered6)
    filtered8 = re.sub("\s\s"," ", filtered7)
    filtered9 = re.sub("\s\s"," ", filtered8)
    filtered10 = re.sub("\s\s"," ", filtered9)
    filtered11 = re.sub("\s\s"," ", filtered10)
    filtered12 = re.sub("\s\s"," ", filtered11)
    filtered13 = re.sub("\s\s"," ", filtered12)
    filtered14 = re.sub("\s\s"," ", filtered13)
    filtered15 = re.sub("\s\n", "\n", filtered14)
    result = re.sub("\s","\n", filtered15)
    result2 = re.sub("\.","\n.\n", result)
    result3 = re.sub("\!","\n!\n", result2)
    result4 = re.sub("\:","\n:\n", result3)
    result5 = re.sub("\?","\n?\n", result4)
    result6 = re.sub("\d","", result5)
with open("intermediate/preprocessed-data-out.txt", "w") as output:
```

```
    output.write(result6)


# Script, das den Output des Tools GermaNER in eine Liste von Wörtern
(Named Entities) konvertiert AP "Identifikation externer Impactrefe-
renzen")
# Autorin: Maria Becker


import fileinput
import re
import sys
with open(sys.argv[1]) as test_text:
    data = test_text.read()
    filtered = re.sub("(\(Hg|\(Hrsg|\(.*\))", "", data)
    filtered2 = re.sub("(\"|\'|-|\(|\))", "", filtered)
    filtered3 = re.sub(",\s*\n","\n", filtered2)
    filtered4 = re.sub("\/","\n", filtered3)
with open("intermediate/test-8.txt", "w") as output:
    output.write(filtered4)


with open('intermediate/test-8.txt') as in_file, open(sys.argv[2],
'w') as out_file:
    for line in in_file:
        if len(line) <= 30:
            out_file.write(line)


# Script, das aus der Liste von Named Entities die Entities extrahiert,
die mehr als einmal vorkommen (AP "Identifikation externer Impactre-
ferenzen")
# Autorin: Maria Becker
```

```
import fileinput
import re
import sys
from collections import Counter
with open(sys.argv[1]) as f:
    c=Counter(c.strip().lower() for c in f if c.strip()) #for case-
insensitive search
    for line in c:
        if c[line]>1:
            print (line)
```


FKZ 01IO2002A
Version 1.1

4.6. Ad AP 4: Statische Auswertung des annotierten Datensatzes

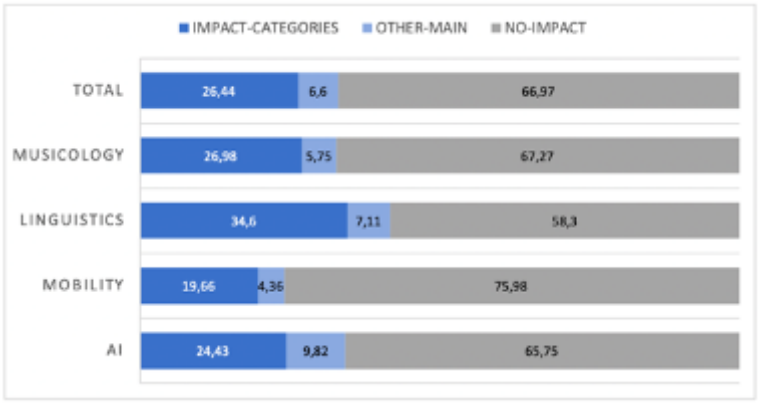

Abbildung 4-21 Anteil impactindizierender Sätze im annotierten Datensatz (in Prozent)

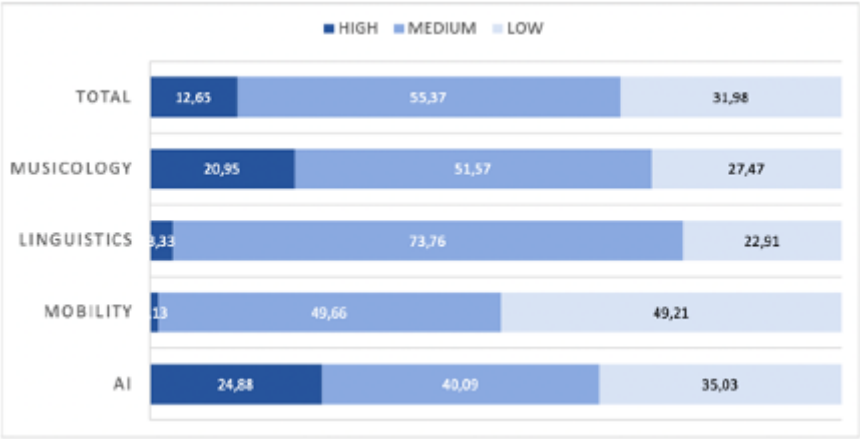

Abbildung 4-22 Impactintensität der impactrelevanten Sätze (in Prozen)

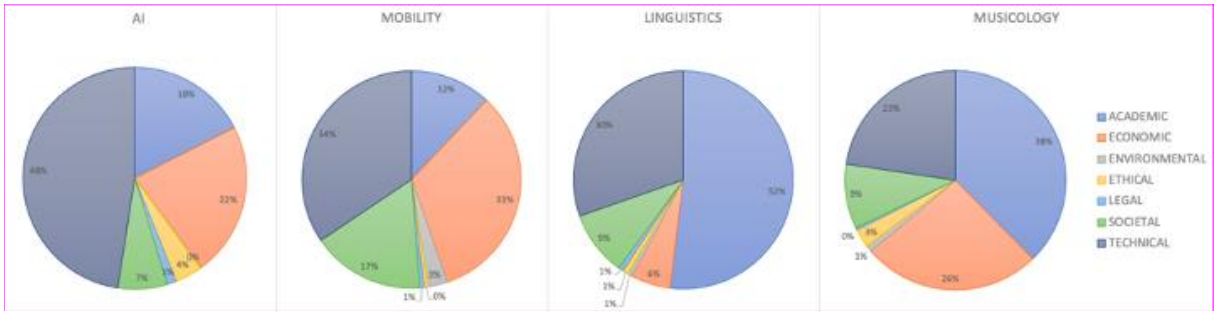

Abbildung 4-23 Verteilung der Impactkategorien auf die Domänen (in Prozent)

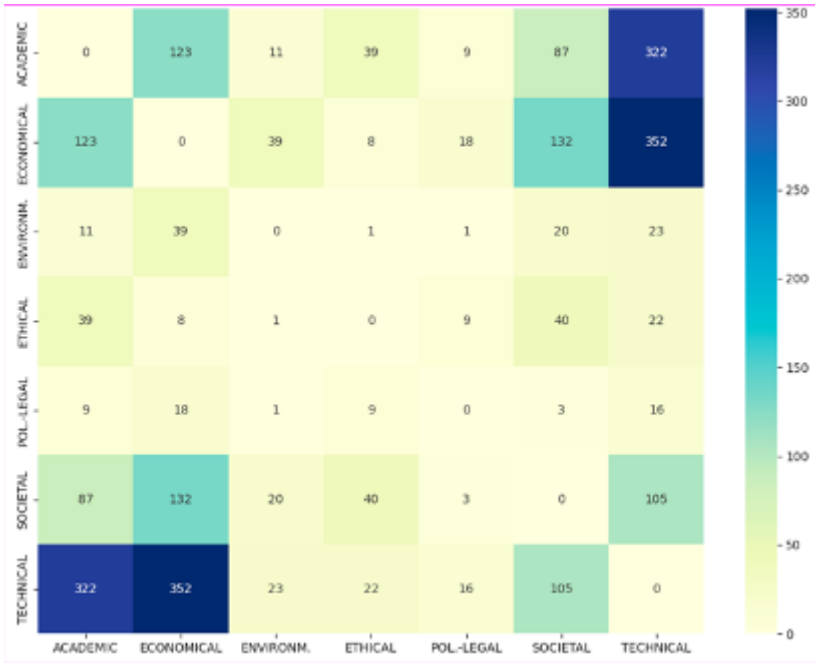

Abbildung 4-24 Symmetrische Heatmap gemeinsam auftretender Impactkatgorien (Multi-Label; absolute Zahlen)