

PsychArchives-ESM 1

Elektronisches Supplement zu Bittermann, Greiner & Fischer: "Unterscheiden sich die Forschungsinteressen von Frauen und Männern in der Psychologie?"

Im Folgenden wird das methodische Vorgehen in Ergänzung zum Methodenabschnitt im Artikel ausführlicher beschrieben.

Software

Die Analysen wurden mit RStudio Version 1.1.456 (RStudio Team, 2016) basierend auf R Version 3.5.1 (R Core Team, 2018) durchgeführt. Für die Aufbereitung der Textdaten wurde *quanteda* 1.4.1 (Benoit et al., 2018), für die Bestimmung des Geschlechts *genderizeR* 2.1.0 (Wais, 2006), für Topic Modeling *topicmodels* 0.2-8 (Grün & Hornik, 2011) und für die Ermittlung der zeitlichen Trends wurden die Pakete *nnet* 7.3-12 (Venables & Ripley, 2002) sowie *quantqual* 0.0.2 (Fischer, 2019) verwendet.

Bestimmung des Geschlechts

Die Datengrundlage bildeten $N = 18\,050$ in der psychologischen Referenzdatenbank PSYNDEX nachgewiesene Dissertationen aus dem deutschen Sprachraum und den Jahren 1968 bis 2017. Dissertationen wurden für die Untersuchung von Forschungsinteressen deshalb ausgewählt, da ihr Inhalt klar der wissenschaftlichen Arbeit einer Person zugeordnet werden kann. Selbst bei kumulativen, publikationsbasierten Promotionen formt die promovierende Person den thematischen Zusammenhang.

Das Geschlecht der verfassenden Person wurde anhand des Vornamens (Naldi, Luzi, Valente & Parenti, 2004) mithilfe des *genderizeR*-Klassifikationsalgorithmus (Wais, 2006) bestimmt. Fell und König (2016) berichten in ihrer szientometrischen Anwendung dieses Algorithmus eine Korrelation von $r = .95$ zwischen wahrem Geschlecht und der automatischen Zuordnung. Wir haben in den 1787 Fällen, in denen der Algorithmus mit einer Sicherheit von weniger als 95 % oder gar nicht zuordnen konnte, sowie bei Unisex-Vornamen (z. B. Dominique, Eike, Luca), das Geschlecht durch manuelle Recherche bestimmt. Insgesamt 79 Personen konnten nach diesem Vorgehen keinem Geschlecht zugeordnet werden, womit der finale Datensatz aus $n = 17\,971$ Dissertationsnachweisen bestand.

Topic Modeling

Controlled Terms

Der Inhalt der Dissertationen wurde wie bei Bittermann und Fischer (2018) anhand der standardisierten Schlagworte in PSYINDEX erfasst (englisch: „Controlled Terms“, im Folgenden: CT). Diese CT sind nicht die freien Schlagworte der Autorinnen und Autoren, sondern sind im *Thesaurus of Psychological Index Terms* (Tuleya, 2007) der American Psychological Association definiert¹. Aus einem Pool von aktuell über 6500 Begriffen werden die CT vom wissenschaftlichen Personal der Informations- und Recherchedienste am ZPID vergeben. Damit wird sichergestellt, dass die zentralen Inhalte einer Publikation durch diese CT abgebildet sind (Gerards, Gerards, Kuhberg-Lasson, Singleton & Trierweiler, 2014, S. 13). Die CT korrespondieren somit mit den Inhalten der Abstracts. Daher repräsentieren CT genau das, was Text Mining von Freitexten zum Ziel hat: Die Extraktion bedeutsamer Information aus Texten (Allahyari et al., 2017). Die Verwendung von CT löst außerdem viele Probleme der Datenaufbereitung von Freitexten (etwa Synonyme, Flexionen, sog. *Stop Words*² usw.) und bietet bei einem Korpus, das die thematische Breite der gesamten Psychologie abdeckt³, einen optimalen Kompromiss aus Detailgrad und Zusammenfassung von Inhalten (für weitere Vorteile siehe Bittermann & Fischer, 2018, S. 5 f).

Von den Vorteilen der CT für unsere Fragestellung abgesehen hätte die Verwendung von Abstracts Mehrsprachigkeit zum Problem gehabt. Mehrsprachigkeit kann bei Topic Modeling zu verschiedenen Themen trotz gleicher Inhalte führen (z. B. englisches "Therapy"-Thema und deutsches "Therapie"-Thema), bzw. die semantische Kohärenz der Themen (Mimno, Wallach, Talley, Leenders & McCallum, 2011) senken (etwa, wenn bei kumulativen Dissertationen in einem deutschsprachigen Abstract die englischen Titel der Einzelbeiträge aufgeführt sind). Daher wird für Topic Modeling in der Regel ein sprachlich homogenes Korpus verwendet, was in unserem Fall zum Ausschluss von 10.3 % der Dissertationen geführt hätte (weder ein deutsches Abstract noch eine Übersetzung wurden publiziert). Dabei ist erstens festzustellen, dass der englischsprachige Ausschluss eine andere thematische Verteilung aufweist als der Rest, vor allem höhere Anteile von Dissertationen aus den Feldern der Neuropsychologie, Elektrophysiologie und Allgemeinen Psychologie⁴. Zweitens ist der Frauenanteil im Ausschluss höher: 65 % vs. 52 % im Rest⁵. Schließlich ist drittens entsprechend des zunehmenden Trends englischsprachiger Publikationen aus den deutschsprachigen Ländern (Schui & Krampen, 2016) zu betonen, dass bei Ausschluss der Dissertationen mit nur englischsprachigen Abstracts ab Publikationsjahr 2001 für manche

¹ Vor allem bei methodischen Arbeiten ist hier sichergestellt, dass ein Verfahren wie etwa "Regressionsanalyse" nicht nur in einer empirischen Analyse eingesetzt wurde, sondern die Methode selbst im Fokus der Publikation stand. Der Thesaurus kann bei ZPID (2016) eingesehen werden.

² Dies sind Wörter, die wenig inhaltliche Relevanz haben (z. B. "die", "ist", "und").

³ Abstracts oder Volltexte sind CT natürlich dann vorzuziehen, wenn das Korpus bereits thematisch eingeschränkt ist und die Detailtiefe von CT nicht mehr ausreichend ist (siehe z. B. Bittermann & Klos, 2019).

⁴ thematische Grobanalyse anhand des PSYINDEX-Klassifikationssystems (ZPID, 2016); χ^2 (93, 28074) = 1808.98, $p < .0001$, Cramer's $V = .25$

⁵ χ^2 (1, 17971) = 105.12, $p < .0001$, $\Phi = .08$

Jahre bis zu 51 % der Arbeiten in unserer Analyse gefehlt hätten⁶. Somit wäre der Ausschluss keine zufällige Stichprobe, vielmehr wären unsere Ergebnisse thematisch und zeitlich verzerrt gewesen.

Topic-Modeling-Verfahren für mehrsprachige Textquellen (z. B. Mimno, Wallach, Naradowsky, Smith & McCallum, 2009; Vulić, De Smet, Tang, & Moens, 2015) erzielen nach unseren Erfahrungen keine befriedigenden Ergebnisse und stellen hohe Anforderungen der Datenvorbereitung, die in unserem Fall nicht erfüllt werden konnten: Thematisch passende Dokumente müssen sprachübergreifend vorab gepaart werden, was bei der sehr ungleichen Verteilung in unserem Korpus nicht adäquat möglich gewesen wäre.

Latent Dirichlet Allocation

Um die den Dissertationen zugrunde liegenden Themen zu bestimmen, wurde Topic Modeling basierend auf *Latent Dirichlet Allocation* (LDA; Blei, Ng & Jordan, 2003) mit Gibbs-Sampling⁷ angewandt. Ziel dieses Verfahrens aus dem maschinellen Lernen ist es, solche Gruppen von Wörtern zu identifizieren, die häufig miteinander in einem Dokument enthalten sind und damit die Themen im Korpus repräsentieren. Zur Identifikation von Themen der Fachliteratur erwies sich Topic Modeling der Verwendung eines Klassifikationssystems (wie z. B. bei König et al., 2015; Preckel & Krampen, 2016) hinsichtlich Detailgrad und Flexibilität überlegen (Bittermann & Fischer, 2018; Ding, 2011), da anhand der Klassifikationskategorien von Literaturdatenbanken nur solche Themen gefunden werden können, die vorab festgelegt wurden.

Bei der Bestimmung der Themen folgten wir den Best-Practice-Empfehlungen von Maier et al. (2018). Es wurden Modelle mit verschiedenen LDA-Hyperparametern⁸ und unterschiedlichen Anzahlen von Themen (75 bis 200) geprüft. Das finale Modell ($k = 100$, $\alpha = 0.01$, $\delta = 0.01$) wurde unter Einbezug von Evaluationsmetriken, anhand qualitativer Prüfung von Dokument-Thema-Zuordnungen und angesichts der besten intersubjektiven Interpretierbarkeit gewählt. Die Themen dieses finalen Modells wurden anschließend hinsichtlich ihrer Reliabilität untersucht (Niekler, 2016). Das heißt, es wurde geprüft, ob sie bei wiederholten Modellinferenzen reproduzierbar sind. Dieses Vorgehen stellt sicher, dass die gefundenen Themen nicht nur aufgrund lokaler Optima der LDA-Modellinferenz zustande kamen. Entsprechend des Ansatzes von Niekler (2016) wurde die Themen-Reliabilität mittels Kosinus-Ähnlichkeit analoger Themen-Paare von mindestens .8 bei acht von zehn wiederholten Modellinferenzen bestimmt. Dieses Vorgehen verringert die Anzahl der Themen zugunsten robusterer Ergebnisse. Für weitere Details verweisen wir auf den kommentierten Code, der in PsychArchives-ESM 2 einsehbar ist.

⁶ $\chi^2 (16, 8846) = 776.88, p < .0001$, Cramer's $V = .30$

⁷ Mit default-Kontrollparametern des *topicmodels*-Pakets; allerdings wurde das beste Modell aus zehn Random Starts mit unterschiedlichen Seeds über das Produkt aus Semantischer Kohärenz (Mimno et al., 2011) und Exklusivität (Roberts et al., 2014) bestimmt, nicht über die heldout-Likelihood.

⁸ alpha-Werte von 0.0001, 0.001 und 0.01, fixed delta = 0.01

Einordnung von Prävalenzdifferenzen

Die Prävalenz beschreibt die durchschnittliche Wahrscheinlichkeit eines Themas im Korpus und wird über die mittlere Dokument-Thema-Wahrscheinlichkeit berechnet. Bei Gleichverteilung aller Themen hat jedes Thema eine Wahrscheinlichkeit von $1/k$, da sich die Wahrscheinlichkeitswerte aller Themen zu 1 addieren. So beträgt bei $k = 100$ Themen die durchschnittliche Prävalenz $1/k = 1/100 = 1\%$ (wenn also jedes Dokument jedes Thema gleich stark behandelt). Dies bedeutet also, dass die beobachteten Prävalenzen der einzelnen Themen umso kleiner ausfallen, je mehr Themen im Modell enthalten sind. Gleiches gilt entsprechend für die Prävalenzdifferenzen zweier Themen oder zweier Gruppen.

Um nun zu beurteilen, wann eine absolute Prävalenzdifferenz ($|D|$) als groß oder klein angenommen werden kann, ist $1/k$ ein hilfreicher Ausgangspunkt. Aufgrund fehlender Standards in der Topic-Modeling-Literatur haben wir folgendes Kriterium festgelegt: Eine Prävalenzerhöhung, die die Prävalenz eines anderen Themas – unter Konstanzhaltung aller anderen Themenprävalenzen – mehr als halbiert (also die halbe durchschnittliche Prävalenz $= 0.5 \cdot 1/k$), wurde als ein größerer Unterschied angesehen. Differenzen $\leq 0.5 \cdot 1/k$ erachten wir entsprechend als einen geringen Unterschied. Im Falle von $k = 100$ Themen bedeutet dies, dass $|D| = 0.5\%$ der Cut-Off zwischen größeren und geringen Unterschieden ist. Dieser Cut-Off gilt auch nach Ausschluss unreliabler Themen (s. o.), da die Prävalenzen der eingeschlossenen Themen während der Inferenz eines Modells mit $k = 100$ Themen berechnet wurden.

Frauenanteile

Die Frauenanteile an den Themen können nicht unmittelbar aus dem Topic Model abgeleitet werden. Dazu muss erst festgelegt werden, wann ein Dokument zu einem Thema gezählt werden kann. Wir haben dies wie folgt operationalisiert: Der Frauenanteil wurde anhand der Anzahl derjenigen Dissertationen bestimmt, in denen das jeweilige Thema das insgesamt Dominierende ist (d. h., der Inhalt der Dissertation lässt sich zu mindestens 50 % diesem Thema zuordnen und zu höchstens 50 % anderen Themen⁹). Wenn also etwa von 100 Dissertationen 40 eine Dokument-Thema-Wahrscheinlichkeit von mindestens 50 % aufweisen, dann werden diese 40 Dissertationen als zu diesem Thema zugehörig gezählt. Anschließend wird der Frauenanteil an diesen 40 Dissertationen ausgezählt. Zwar geht durch diesen Cut-Off Information verloren, für illustrative Zwecke ist der Frauenanteil an einem Thema allerdings intuitiver erfassbar als Prävalenzdifferenzen.

⁹ z. B.: Die Behandlung der Depression ist Hauptthema der Dissertation, thematisch werden in geringerem Umfang auch Diagnostik und Epidemiologie angeschnitten.

Zeitliche Trends

Um die Zeitverläufe der Themen von Schwankungen zu befreien, wurden Trendkurven getrennt nach Geschlecht mithilfe von *Multilayer Perceptrons* (MLP) berechnet. Die MLPs ermöglichen nichtlineare Regressionsfunktionen mit einer minimalen Summe von quadratischen Residuen für jedes Thema. Für jedes MLP wurde eine Struktur mit zwei Hidden Units berechnet¹⁰, um die durchschnittliche Themenwahrscheinlichkeit als nichtlineare Funktion des Erscheinungsjahres zu modellieren (vgl. Bittermann & Fischer, 2018). Zwei Hidden Units wurden gewählt, um einerseits – im Gegensatz zu weniger komplexen Strukturen – auch nicht-monotone Funktionen zu ermöglichen und andererseits – im Vergleich zu komplexeren Strukturen – das Risiko (und das Ausmaß) einer Überanpassung (engl. *Overfitting*) des Modells an die Daten zu reduzieren (Fischer, 2015). Die zeitlichen Trends wurden schließlich auf drei Dimensionen untersucht:

- (1) Dimension “Geschlecht”, das heißt die mittlere Differenz zwischen der weiblichen und männlichen Trendkurve,
- (2) Dimension “Konvergenz” (im Sinne einer über die Zeit zunehmenden Nivellierung zweier Trends) anhand eines lineares Regressionsmodells mit Prädiktor Publikationsjahr und Kriterium Differenz der weiblichen und männlichen Trendkurve, sowie
- (3) Dimension “Trend”, das heißt der lineare Gesamttrend des Themas, unabhängig vom Geschlecht.

Liste der verwendeten R-Pakete

```
> sessionInfo()
R version 3.5.1 (2018-07-02)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

Matrix products: default

locale:
[1] LC_COLLATE=German_Germany.1252 LC_CTYPE=German_Germany.1252    LC_MONETARY=German_Germany.1252
LC_NUMERIC=C
[5] LC_TIME=German_Germany.1252

attached base packages:
[1] stats    graphics grDevices utils      datasets methods  base

other attached packages:
[1] nnet_7.3-12      quantqual_0.0.2  genderizeR_2.1.0 data.table_1.12.0 stringr_1.4.0      dplyr_0.8.0.1
topicmodels_0.2-8
[8] quanteda_1.4.1

loaded via a namespace (and not attached):
[1] modeltools_0.2-22 tidyselect_0.2.5  NLP_0.2-0          slam_0.1-45        reshape2_1.4.3     purrr_0.3.1
```

¹⁰ Um im Rahmen des Schätzprozesses lokale Minima zu vermeiden, wurde analog zu Fischer (2015) der Schätzprozess zehn mal wiederholt und das beste Modell zur Analyse der jeweiligen Trendkurve herangezogen.

[7] splines_3.5.1	lattice_0.20-38	colorspace_1.4-0	generics_0.0.2	stats4_3.5.1	yaml_2.2.0
[13] survival_2.43-3	proclim_2018.04.18	rlang_0.3.1	ModelMetrics_1.2.2	pillar_1.3.1	glue_1.3.1
[19] withr_2.1.2	foreach_1.4.4	plyr_1.8.4	lava_1.6.5	timeDate_3043.102	munsell_0.5.0
[25] gtable_0.2.0	recipes_0.1.4	codetools_0.2-16	caret_6.0-81	parallel_3.5.1	tm_0.7-6
[31] class_7.3-15	Rcpp_1.0.0	spacyr_1.0	scales_1.0.0	ipred_0.9-8	RcppParallel_4.4.2
[37] fastmatch_1.1-0	stopwords_0.9.0	ggplot2_3.1.0	stringi_1.3.1	grid_3.5.1	tools_3.5.1
[43] magrittr_1.5	lazyeval_0.2.1	tibble_2.0.1	crayon_1.3.4	pkgconfig_2.0.2	MASS_7.3-51.1
[49] Matrix_1.2-15	xml2_1.2.0	lubridate_1.7.4	gower_0.2.0	assertthat_0.2.0	rstudioapi_0.9.0
[55] iterators_1.0.10	R6_2.4.0	rpart_4.1-13	nlme_3.1-137	compiler_3.5.1	

Literatur

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint*. <https://arxiv.org/abs/1707.02919v2>

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3 (30), 774. <http://doi.org/10.21105/joss.00774>

Bittermann, A. & Fischer, A. (2018). How to identify hot topics in psychology using topic modeling. *Zeitschrift für Psychologie*, 226 (1), 3–13. <https://doi.org/10.1027/2151-2604/a000318>

Bittermann, A. & Klos, E. M. (2019). Ist die psychologische Forschung durchlässig für aktuelle gesellschaftliche Themen? Eine szientometrische Analyse am Beispiel Flucht und Migration mithilfe von Topic Modeling. *Psychologische Rundschau*, 70 (4), 239–249. <https://doi.org/10.1026/0033-3042/a000426>

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://doi.org/10.1162/jmlr.2003.3.4-5.993>

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, 5 (1), 187–203. <https://doi.org/10.1016/j.joi.2010.10.008>

Fell, C. B. & König, C. J. (2016). Is there a gender difference in scientific collaboration? A scientometric examination of co-authorships among industrial–organizational psychologists. *Scientometrics*, 108 (1), 113–141. <https://doi.org/10.1007/s11192-016-1967-5>

Fischer, A. (2015). How to determine the unique contributions of input-variables to the nonlinear regression function of a multilayer perceptron. *Ecological Modelling*, 309, 60–63. <https://doi.org/10.1016/j.ecolmodel.2015.04.015>

- Fischer, A. (2019). *quantqual: Software package for analyzing quantitative and qualitative data*. [Computer software]. Verfügbar unter <https://github.com/AndreasFischer1985/quantqual>
- Gerards, A., Gerards, M., Kuhberg-Lasson, V., Singleton, K. & Trierweiler, L. (2014). Dokumentationsmanual PSYNDEX (aktualisierte Ausgabe). Trier: ZPID. Verfügbar unter <https://www.psyndex.de/download/PSYNDEXmaterial/psyndex-dokumentationsmanual.pdf>
- Grün, B. & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40, 1–30. <https://doi.org/10.18637/jss.v040.i13>
- Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri H. & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12 (2–3), 93–118. <https://doi.org/10.1080/19312458.2018.1430754>
- Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009, August). Polylingual topic models. In P. Koehn & R. Mihalcea (Hrsg.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2* (S. 880–889). Stroudsburg, PA: Association for Computational Linguistics. <https://www.aclweb.org/anthology/D09-1092.pdf>
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (262–272). Association for Computational Linguistics. Chicago.
- Naldi, F., Luzi, D., Valente, A., & Parenti, I. V. (2004). Scientific and technological performance by gender. In H. F. Moed, W. Glänzel, & U. Schmoch (Hrsg.), *Handbook of quantitative science and technology research* (S. 299–314). Dordrecht: Kluwer. <http://dx.doi.org/10.1007/1-4020-2755-9>
- Niekler, A. (2016). *Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen*. Dissertation, Universität Leipzig. Verfügbar unter: http://asv.informatik.uni-leipzig.de/publication/file/350/Niekler_Diss.pdf
- Preckel, F. & Krampen, G. (2016). Entwicklung und Schwerpunkte in der psychologischen Hochbegabungsforschung. *Psychologische Rundschau*, 67 (1), 1–14. <https://doi.org/10.1026/0033-3042/a000289>
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. [Computer software]. Verfügbar unter: <https://www.R-project.org/>

- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58 (4), 1064–1082. <https://doi.org/10.1111/ajps.12103>
- RStudio Team. (2016). *RStudio: Integrated development for R* [Computer software]. Boston, MA: RStudio, Inc. Verfügbar unter: <http://www.rstudio.com/>
- Schui, G. & Krampen, G. (2016). ZPID-Monitor 2014 zur Internationalität der Psychologie aus dem deutschsprachigen Bereich: Der ausführliche Bericht. *ZPID Science Information Online*, 16 (2). Verfügbar unter https://www.zpid.de/pub/research/zpid-monitor_2014_lang.pdf
- Tuleya L. G. (Hrsg.). (2007). *Thesaurus of psychological index terms* (11th ed.). Washington, DC: American Psychological Association.
- Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth Edition). Springer, New York.
- Vulić, I., De Smet, W., Tang, J., & Moens, M. F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51, 111–147. <https://doi.org/10.1016/j.ipm.2014.08.003>
- Wais, K. (2006). Gender Prediction Methods Based on First Names with genderizeR. *The R Journal*, 8 (1), 17–37. <http://doi.org/10.32614/RJ-2016-002>
- ZPID – Leibniz-Zentrum für Psychologische Information und Dokumentation (Hrsg.). (2016). *PSYINDEX Terms* (10. Aufl.). Trier: ZPID. Verfügbar unter: <https://www.psyindex.de/pub/info/PSYINDEXterms2016.pdf>