# Eine quantitative Analyse anhand von Big Data aus dem Katalog:
# Wohin entwickelt sich die Romanistik?

## *Fachspezifische Angebote der SUB Göttingen - von Anglistik bis Zentralasienkunde, 5.5.2022*

---


**Dr. José Calvo Tello**

UNIVERSITÄT GÖTTINGEN

NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN | SUB

- *Digital Humanities*? Ok
- *Big Data* für die Geisteswissenschaften?


ME WHEN SOMEBODY TALKS ABOUT
BIG DATA APPLIED TO HUMANITIES
imgflip.com

# Katalogdaten als Forschungsobjekt

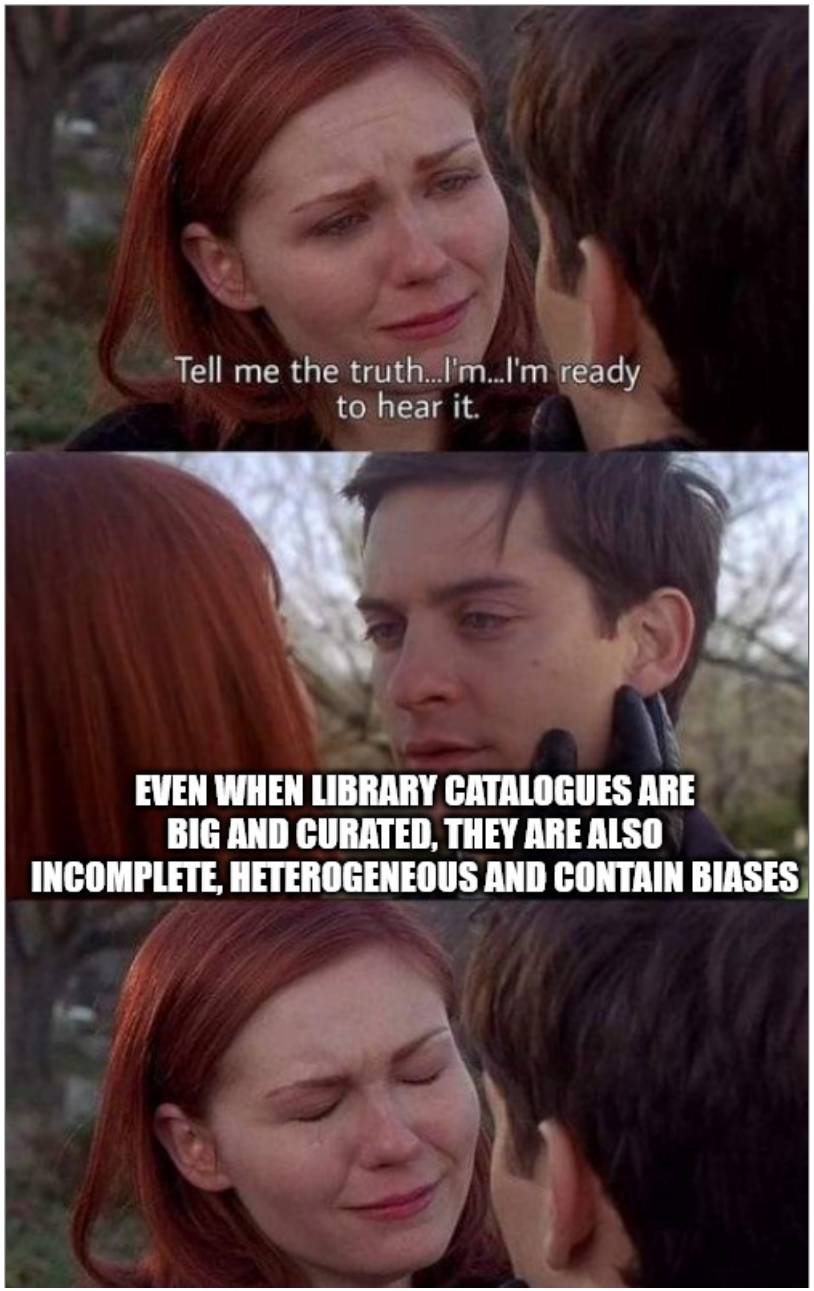
| | |
|---|---|
| **Titel:** | Horses and riding equipment in Indian art / Jean Deloche |
| **VerfasserIn:** | Deloche, Jean |
| **Ausgabe:** | English ed |
| **Sprache/n:** | Englisch |
| **Sprache des Originals:** | Franzoesisch |
| **Veröffentlichungsangabe:** | Madras : Indian Heritage Trust, 1990 |
| **Umfang:** | 82 S : zahlr. Ill ; 30 cm |
| **Einheitssachtitel:** | Le cheval et son harnachement dans l'art indien <engl.> |
| **Anmerkung:** | Includes bibliographical references |
| **ISBN:** | : Rs150.00 |
| **Schlagwörter:** | *Indien / Kunst / Pferd<br>*Horses in art ; Horses -- Equipment and supplies -- Pictorial works ; Art, Indic |
| **Sachgebiete:** | 20.41 Asiatische Kunst |
| **Mehr zum Thema:** | Klassifikation der Library of Congress: N7668.H6<br>Dewey Dezimal-Klassifikation: 704.9432 ; 704.9/432<br>Regensburger Verbund-Klassifikation: LH 84680: Kunstgeschichte / Allgemeines. Allgemeine Kunstgeschichte / Kunstgeschichte einzelner Gattungen der Kunst / Ikonographie, Ikonologie / Weltliche Ikonographie / Ikonographie der Tiere (Bestiarien) / Einzelne Tiere (soweit sie nicht ausschließlich als biblische oder christliche Symbole aufgefaßt sind) / Sonstige ; LO 88380: Kunstgeschichte / Kunst nach Ländern bzw. Kontinenten / Asien und Vorderer Orient / Kunstgeschichte Südasiens / Kunstgeschichte Indiens / Indische Kunstgeschichte nach Gattungen / Sonstige Gattungen (u.a. Ornament) |
| **Sachgebiete:** | GED 130 Buddhistische Kunst. Hinduistische Kunst {Geschichte der bildenden Kunst}<br>GEC 700 Profane Ikonographie {Kunst} |
| **Standort:** | SUB Zentralbibliothek |
| **Signatur:** | **A 96 B 35008** |
| **Ausleihstatus:** | Ausleihbestand<br>Derzeit verfuegbar Bestellen |

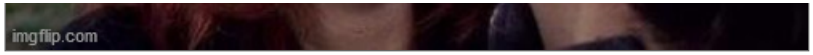
## Katalogdaten als Forschungsobjekt

- Wachsendes Interesse an Daten aus Bibliographien und Katalogen
- *Bibliographic Data Science*
- Vorteile
  - Große Datenmengen (*Big Data*)
  - Kuratierte Daten (in vielen Fällen normiert)

# Big Curated Data!

- Allerdings…


Tell me the truth...I'm...I'm ready to hear it.
EVEN WHEN LIBRARY CATALOGUES ARE BIG AND CURATED, THEY ARE ALSO INCOMPLETE, HETEROGENEOUS AND CONTAIN BIASES


imgflip.com

## Ziele

- Entwicklung der Romanistik in den letzten 40 Jahren?
- Erwartung für die Romanistik für die nächsten Jahre?

# Wofür brauche ich diese Analyse als Fachreferent?

## Wofür brauche ich das als Fachreferent?

- Um die Angebote der Bibliothek an die aktuellen Entwicklungen der Romanistik anzupassen
  - Bestand
  - Lesesaal
  - Digital vs. Print
  - Beratung

# Daten

# Quellen: Verbundkataloge

- K10plus (GBV und SWB)
- HeBIS (Hessisches BibliotheksInformationsSystem)
- Art der Veröffentlichung: Monographien, Zeitschriften, Serien, E-Books (auch Open Access)
- Fachinformationsdienste (FIDs):
  - Fachinformationsdienst Romanistik
  - Fachinformationsdienst Lateinamerika, Karibik und Latino Studies

## Auswahl der Veröffentlichungen der Romanistik

- Göttinger Online-Klassifikation (GOK)

## Romanistik in Klassifikationssystemen:

- Basisklassifikation (BK)
  - regex = "18.[23]\d"
- Sachgruppe der DNB
  - regex = "^(4[456]\d|8[456])"
- Dewey Decimal Classification (DDC)
  - regex = "^(4[456]\d|8[456])"
- Regensburger Verbundklassifikation (RVK)
  - regex = "^I.*?"
- Library of Congress Classification (LCC)
  - regex = "^p[cq].*?"
- Göttinger Online Klassifikation (GOK)
  - regex = "^IH-Z.*?"

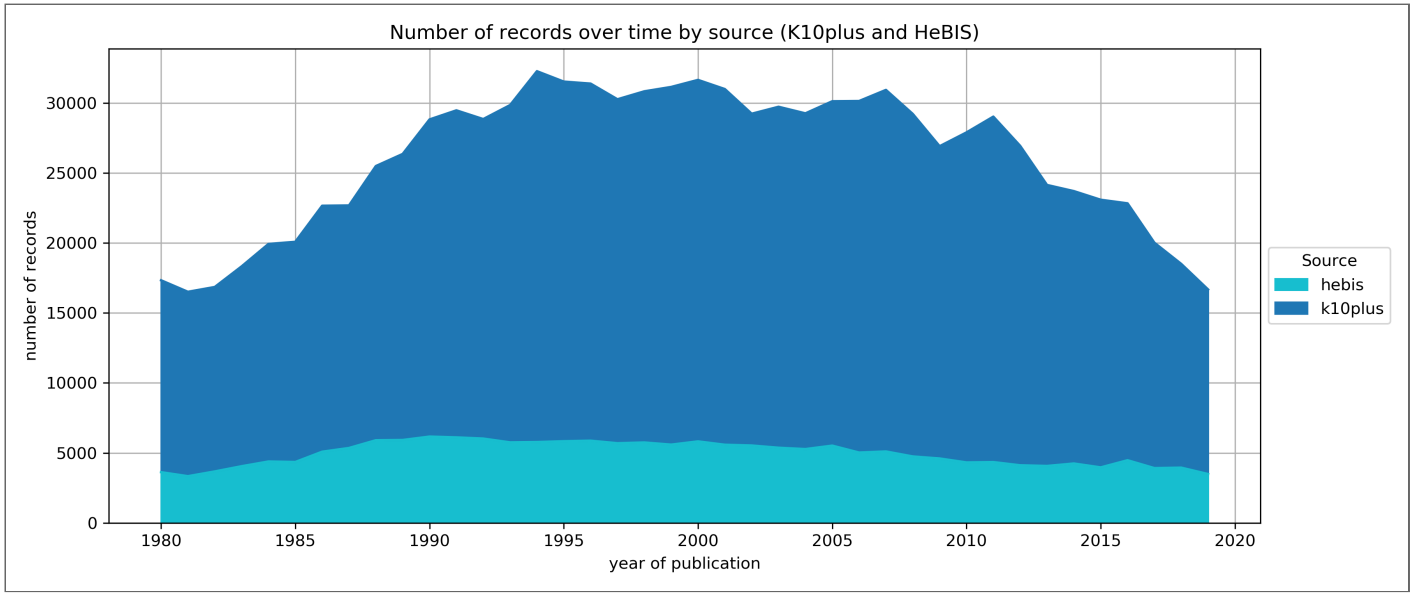
## Sekundärliteratur

- Folgende Klassen wurden ausgeschlossen:
  - Sachgruppen: B
  - BK: 17.97 und 17.98
  - RVK: Klassen mit "Gesammelte Werke" oder "Einzelwerke"
  - Art des Inhalts: Fiktionale Darstellung, Anthologie, Briefsammlung

## Zusammenstellung und Extraktion

1. Zeitliche Grenzen: 1980-2019
2. Daten über Schnittstellen (API) herunterladen
3. Daten extrahieren (xPaths und RegEx) und in eine Tabelle konvertieren
4. Berechnung der relativen Häufigkeit anhand der Anzahl pro Jahr

# Datensatz




- 1.041.404 Titel (unter Berücksichtigung der Anzahl von Bibliotheken, die den Titel im Bestand haben)
- 81% K10plus
- 19% HeBIS
- Daten und Skripte bereits in GitLab veröffentlicht: **https://gitlab.gwdg.de/jose.calvotello/romance-languages-studies-2021**

# Methode

## Lineare Regression

- Bereich des überwachten maschinellen Lernens
- ~~Klassifikation für kategoriale Werte~~
- Ziel ist die Vorhersage eines numerischen Werts

# Seiten: Beispiel

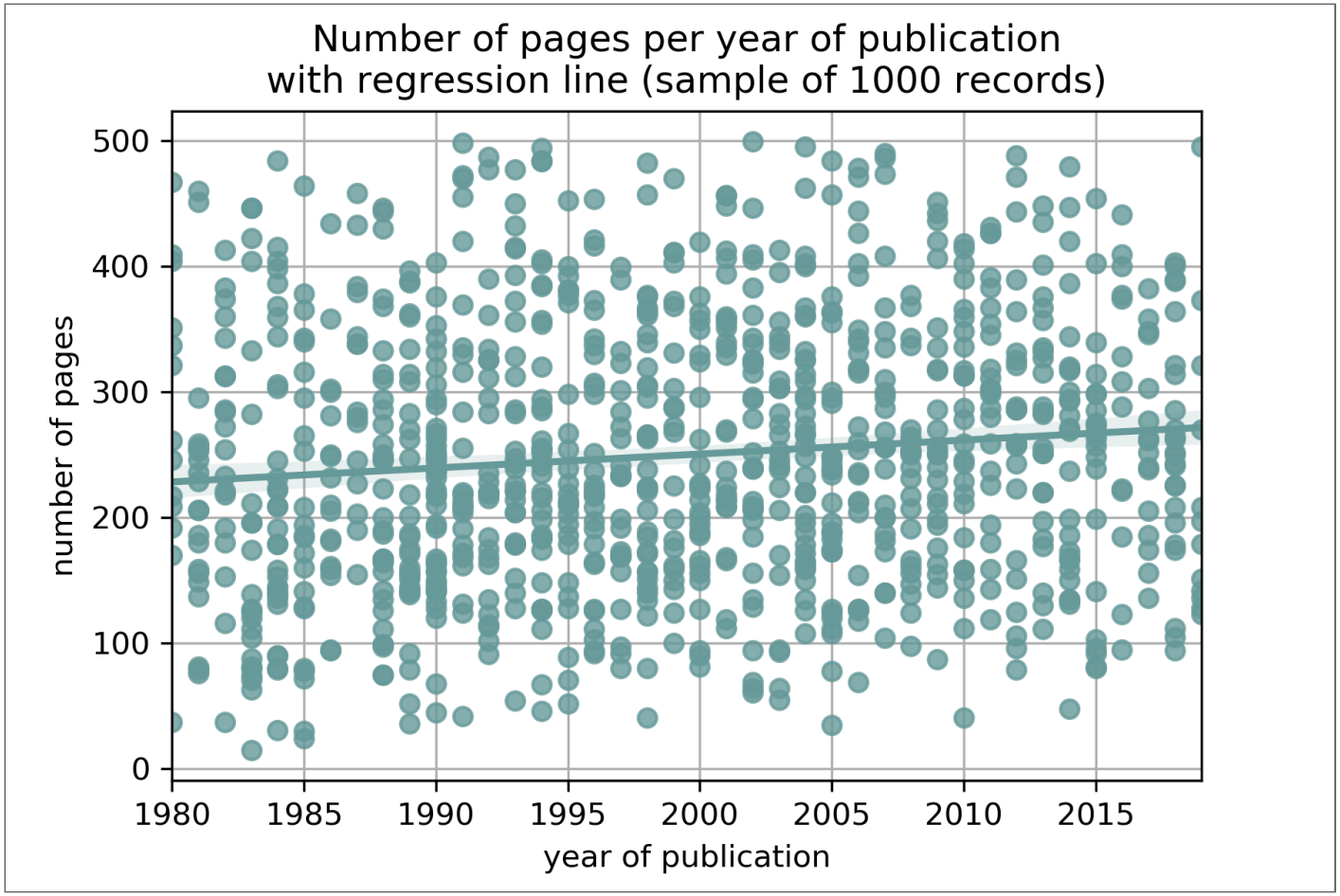
- **Forschungsfrage**: Werden die Veröffentlichungen länger? Nimmt die Anzahl der Seiten im Laufe der Zeit zu?

| | |
|---|---|
| **Titel:** | Horses and riding equipment in Indian art / Jean Deloche |
| **VerfasserIn:** | Deloche, Jean |
| **Ausgabe:** | English ed |
| **Sprache/n:** | Englisch |
| **Sprache des Originals:** | Franzoesisch |
| **Veröffentlichungsangabe:** | Madras : Indian Heritage Trust, 1990 |
| **Umfang:** | 82 S : zahlr. Ill ; 30 cm |
| **Einheitssachtitel:** | Le cheval et son harnachement dans l'art indien <engl.> |
| **Anmerkung:** | Includes bibliographical references |
| **ISBN:** | : Rs150.00 |
| **Schlagwörter:** | *Indien / Kunst / Pferd<br>*Horses in art ; Horses -- Equipment and supplies -- Pictorial works ; Art, Indic |
| **Sachgebiete:** | 20.41 Asiatische Kunst |
| **Mehr zum Thema:** | Klassifikation der Library of Congress: N7668.H6<br>Dewey Dezimal-Klassifikation: 704.9432 ; 704.9/432<br>Regensburger Verbund-Klassifikation: LH 84680: Kunstgeschichte / Allgemeines. Allgemeine Kunstgeschichte / Kunstgeschichte einzelner Gattungen der Kunst / Ikonographie, Ikonologie / Weltliche Ikonographie / Ikonographie der Tiere (Bestiarien) / Einzelne Tiere (soweit sie nicht ausschließlich als biblische oder christliche Symbole aufgefaßt sind) / Sonstige ; LO 88380: Kunstgeschichte / Kunst nach Ländern bzw. Kontinenten / Asien und Vorderer Orient / Kunstgeschichte Südasiens / Kunstgeschichte Indiens / Indische Kunstgeschichte nach Gattungen / Sonstige Gattungen (u.a. Ornament) |
| **Sachgebiete:** | GED 130 Buddhistische Kunst. Hinduistische Kunst {Geschichte der bildenden Kunst}<br>GEC 700 Profane Ikonographie {Kunst} |
| **Standort:** | SUB Zentralbibliothek |
| **Signatur:** | **A 96 B 35008** |
| **Ausleihstatus:** | Ausleihbestand<br>Derzeit verfuegbar ➔ Bestellen |

## Seiten: Fiktionales Beispiel

- 1980: 80 Seiten
- 1981: 81 Seiten
- 1982: 82 Seiten
- 1990: 90 Seiten
- 2020: 120 Seiten
- 2030: ?

# Seiten: Visualisierung




- Durchschnitt von Seiten in **1980: 219**
- Durchschnitt von Seiten in **2019: 256**

## Seiten: Regression

- **slope = 0,9**, p-value < 0,001
- Erwartung:
  - Publikationen werden in 10 Jahren im Durchschnitt 9 Seiten länger sein
  - 2030 Publikationen von 277
  - 2040 Publikationen von 286

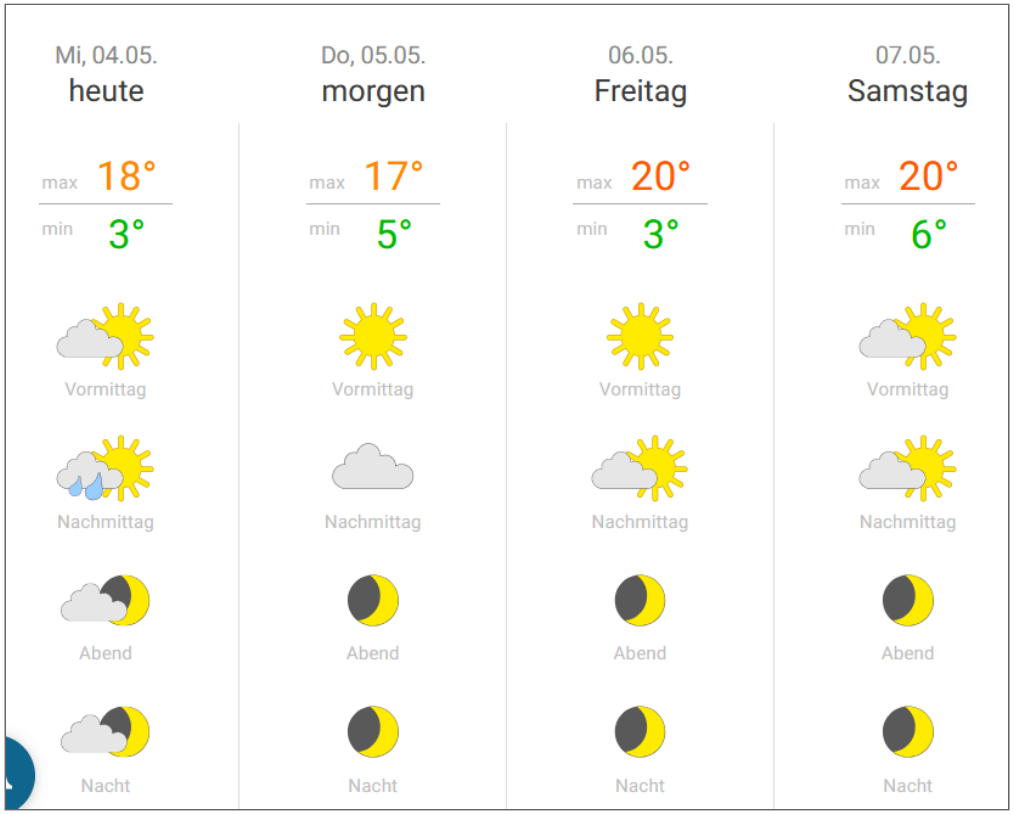
# Aber...


PUBLICATIONS WILL
HAVE 277 PAGES IN 2030
ARE THEY THOUGH?


imgflip.com

## Nein, diese Vorhersagen werden nicht ganz zutreffend

# Wettervorhersagen sind auch nie perfekt, trotzdem nützlich



## Um Entscheidungen zu treffen

- Entweder für den Trend
- Oder dagegen!

## Generelles Problem der Methoden des aktuellen (überwachten) maschinellen Lernens

### Konservativ

- Mit Daten, die wir bereits gesammelt haben, neue Fälle vorhersagen
- Mit Daten der Vergangenheit zukünftige Entwicklungen vorhersagen

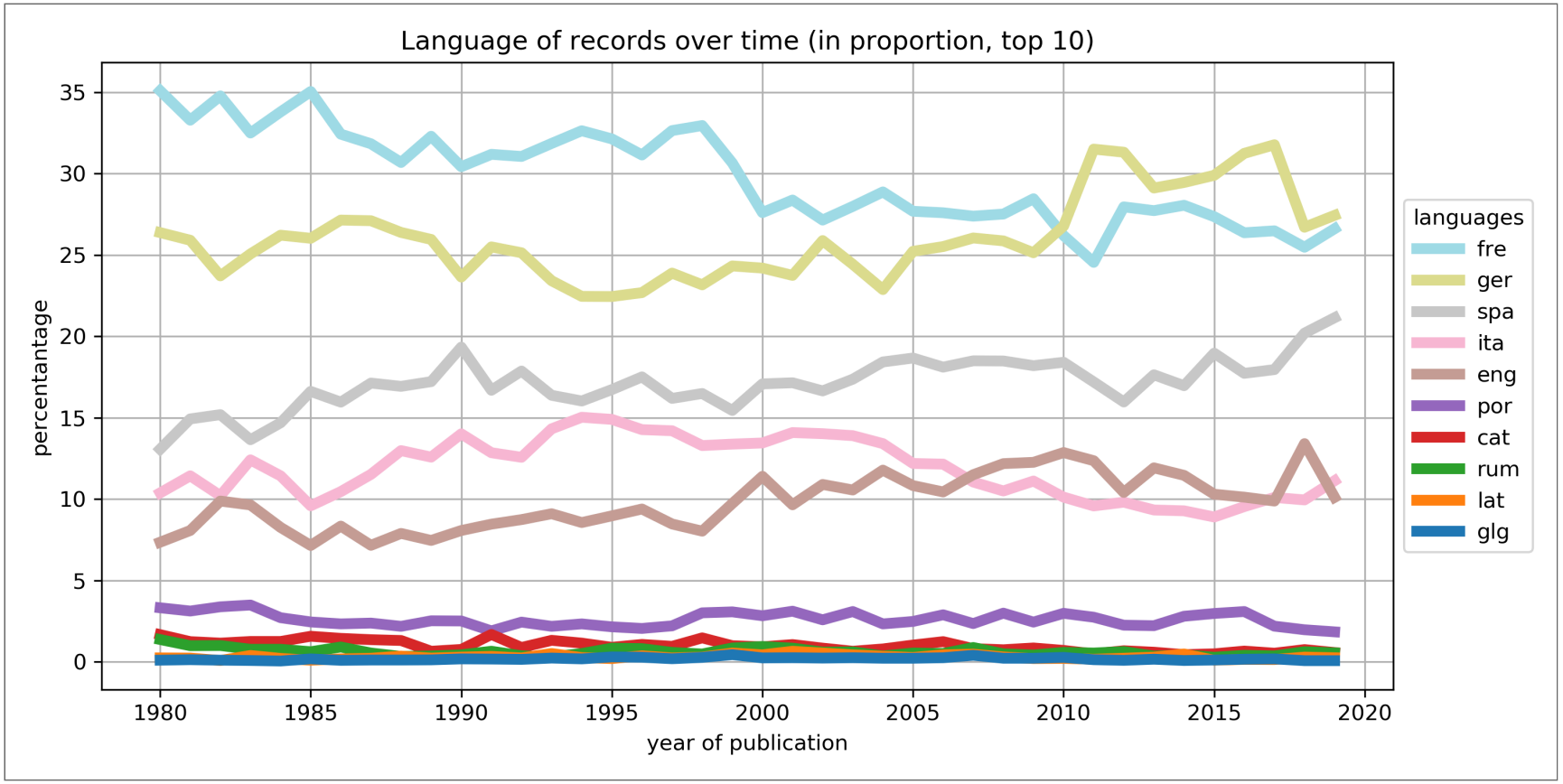
# Wichtig: immer neue kuratierte Daten eintragen

# Kategorien

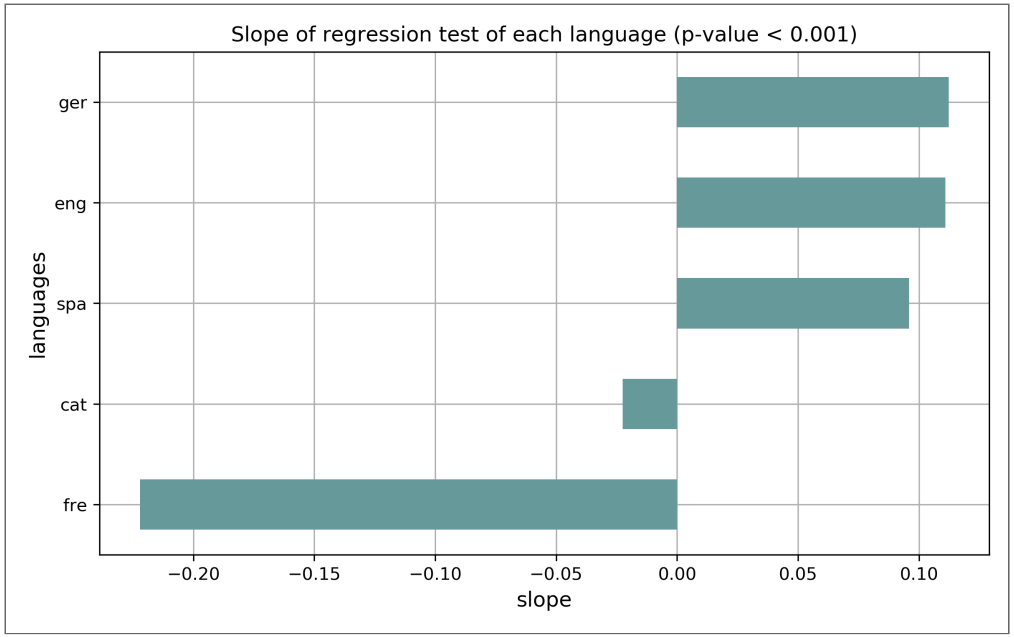
# Sprachen

- Veröffentlichungssprache
- ~~Sprache als Forschungsobjekt~~ (später!)

# Sprachen: Entwicklung



# Sprachen: Slope

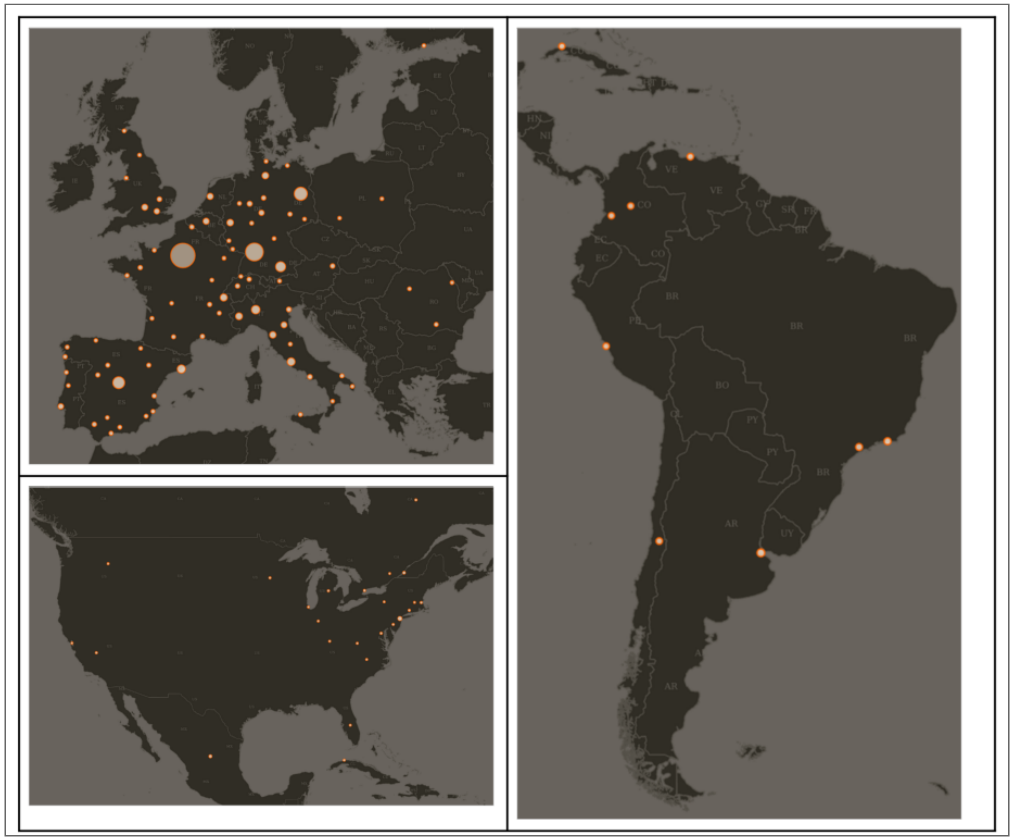



- Erwartung für 2030
  - 29 % Deutsch
  - 24 % Französisch
  - 20 % Spanisch (2. Sprache ab 2040)
  - 13 % Englisch

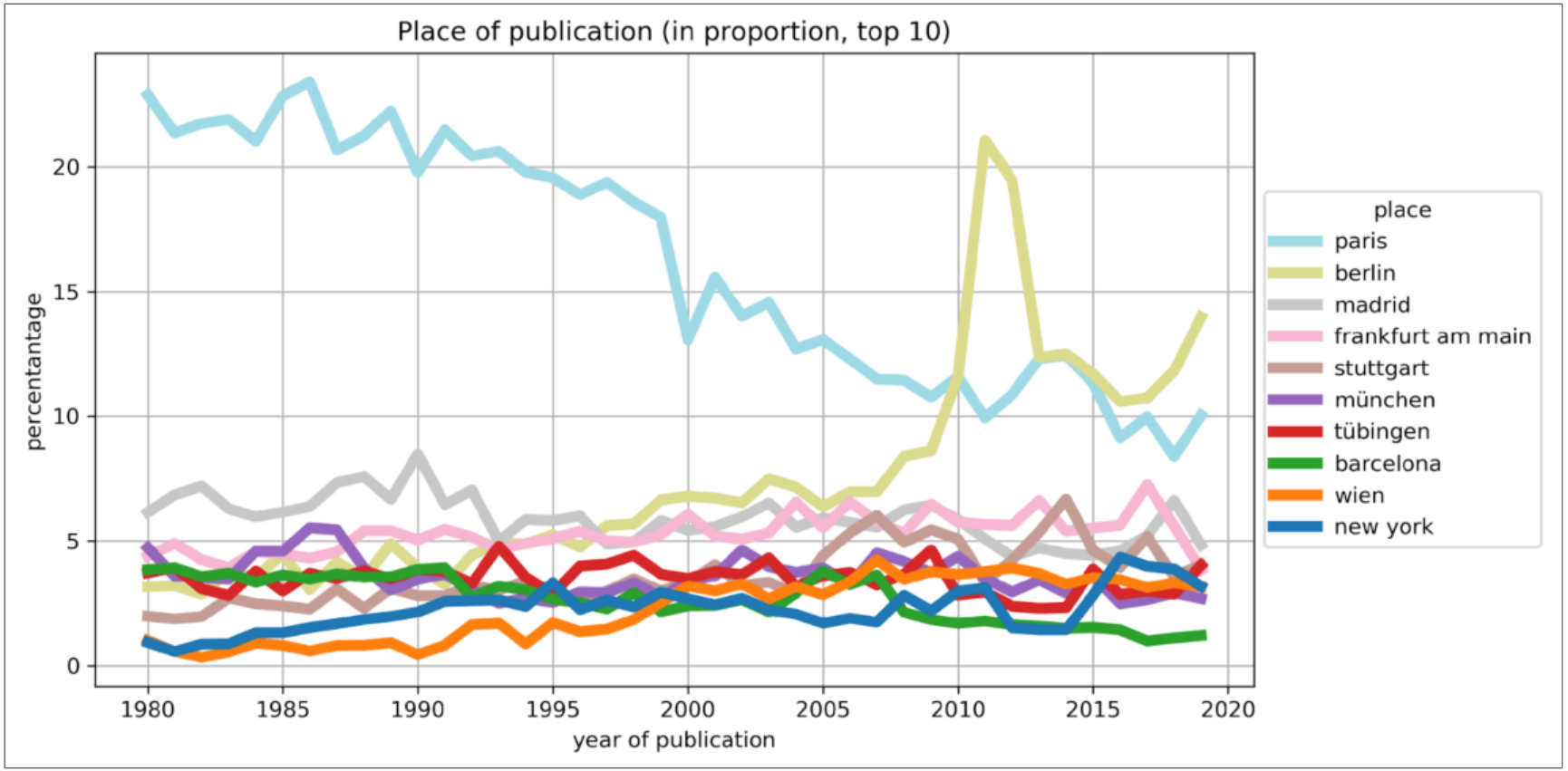
# Ort der Veröffentlichung

# Ort der Veröffentlichung: Visualisierung

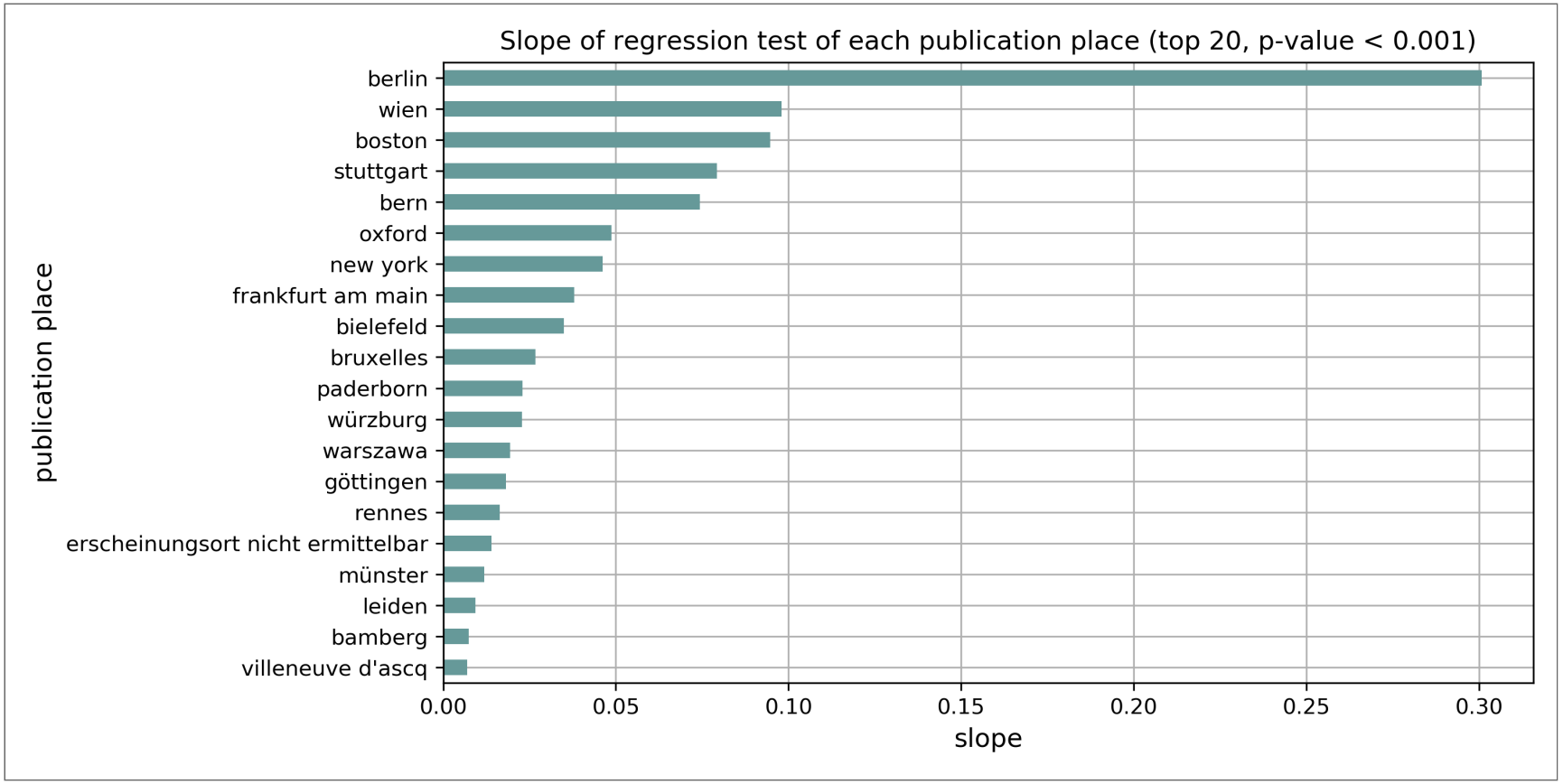
- Sample mit 5.000 Publikationen
- In **DARIAH-DE Geo-Browser visualisiert und veröffentlicht**



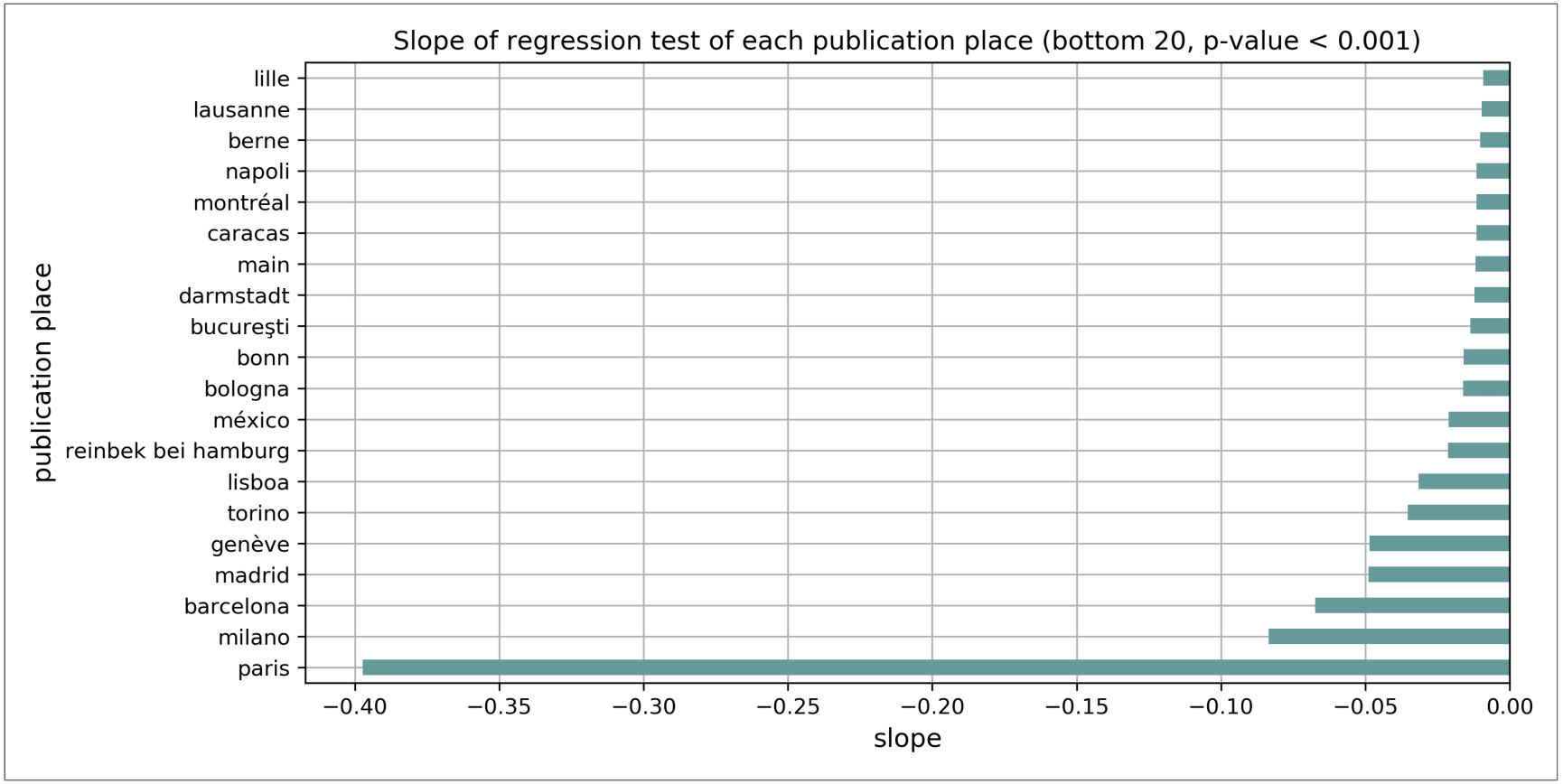
# Ort der Veröffentlichung: Entwicklung



# Ort der Veröffentlichung: Slope (top)

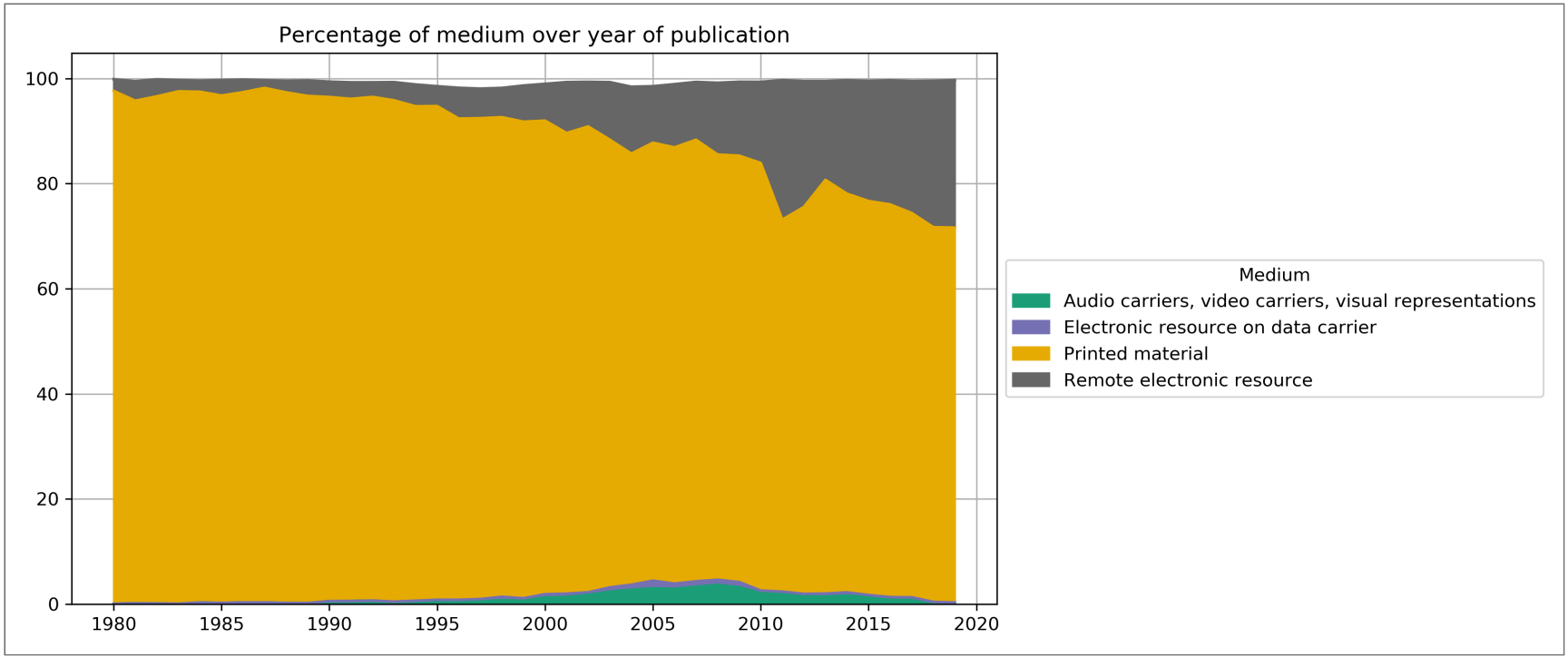


# Ort der Veröffentlichung: Slope (bottom)



# Medium (Print vs. E-Book)

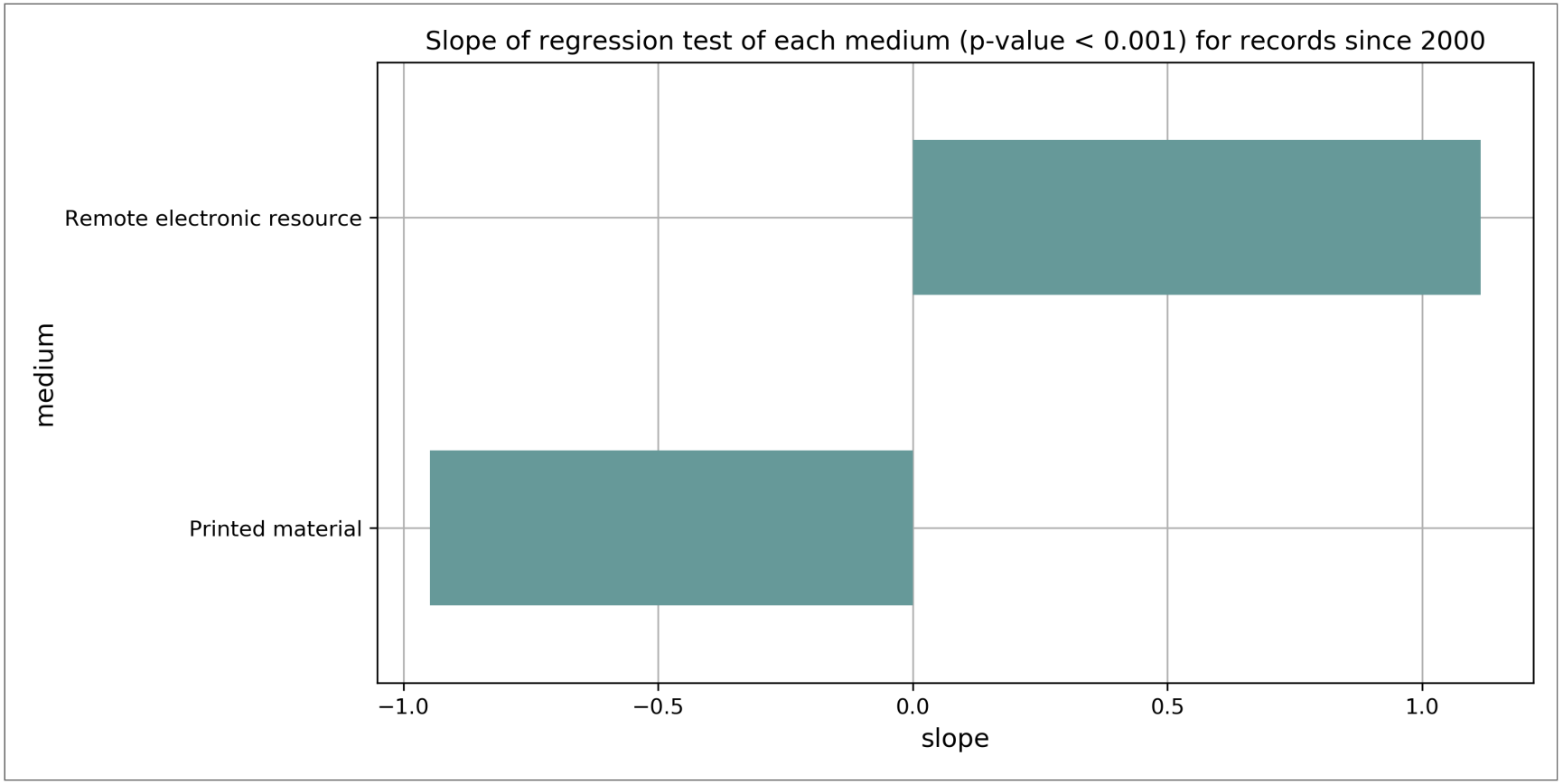
# Medium: Entwicklung



## Medium: Entwicklung

- Gesamte Periode
  - Print: ca. 90 %
  - E-Books: ca. 10 %
- Nach 2010
  - E-Books > 20 %

# Medium: Slope




- Erwartung: 1,1 % mehr E-Books pro Jahr

# Wann werden alle Publikationen der Romanistik als E-Book vorliegen?

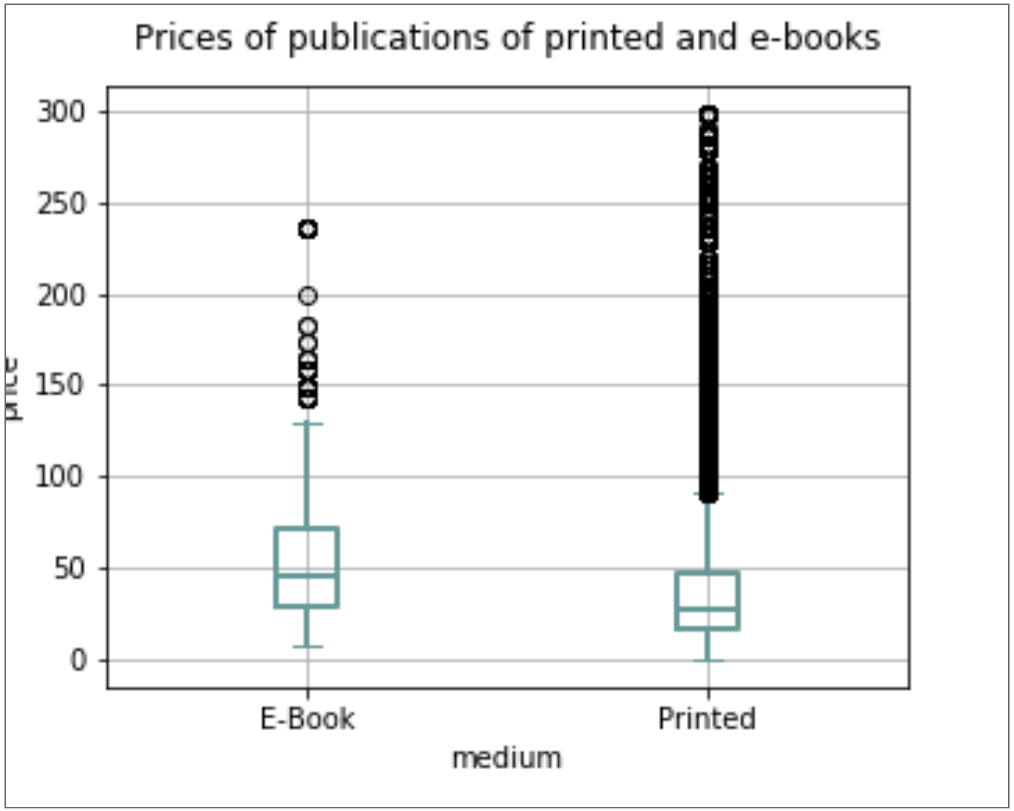
# 100% E-Books in 2080!

## Preis: Print vs. E-Book

- Was ist günstiger? E-Books oder Print?

# Bibliotheken müssen besondere Lizenzen für E-Books erwerben!

# Preis: Print vs. E-Book

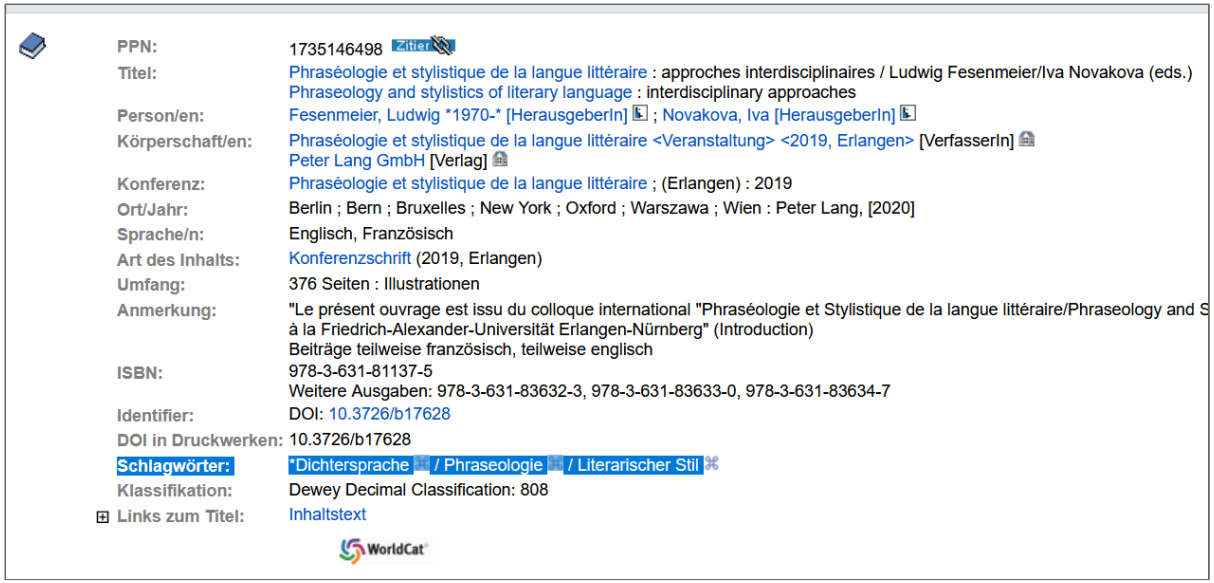



- Print = 28 € Median Preis
- E-Book = 46 € Median Preis
- P-value < 0,001 in Welch’s t-tes
- Die Ergebnisse sind nicht endgültig!

# Themen

# Themen: Schlagwörter

# Schlagwörter im Katalog

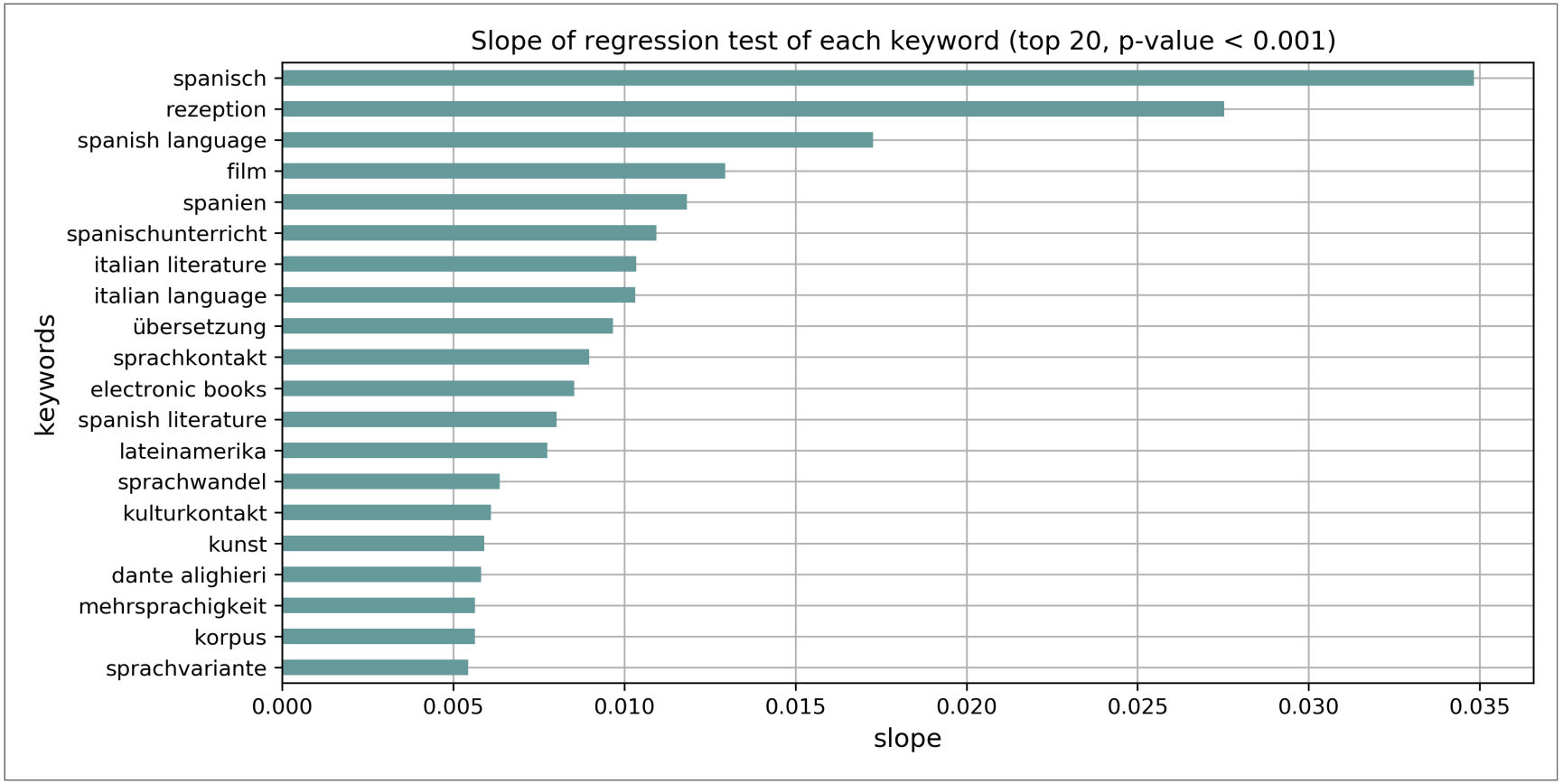


| | |
|---|---|
| PPN: | 1735146498 Zitieren |
| Titel: | Phraséologie et stylistique de la langue littéraire : approches interdisciplinaires / Ludwig Fesenmeier/Iva Novakova (eds.)<br>Phraseology and stylistics of literary language : interdisciplinary approaches |
| Person/en: | Fesenmeier, Ludwig *1970-* [HerausgeberIn] ; Novakova, Iva [HerausgeberIn] |
| Körperschaft/en: | Phraséologie et stylistique de la langue littéraire <Veranstaltung> <2019, Erlangen> [VerfasserIn]<br>Peter Lang GmbH [Verlag] |
| Konferenz: | Phraséologie et stylistique de la langue littéraire ; (Erlangen) : 2019 |
| Ort/Jahr: | Berlin ; Bern ; Bruxelles ; New York ; Oxford ; Warszawa ; Wien : Peter Lang, [2020] |
| Sprache/n: | Englisch, Französisch |
| Art des Inhalts: | Konferenzschrift (2019, Erlangen) |
| Umfang: | 376 Seiten : Illustrationen |
| Anmerkung: | "Le présent ouvrage est issu du colloque international "Phraséologie et Stylistique de la langue littéraire/Phraseology and S<br>à la Friedrich-Alexander-Universität Erlangen-Nürnberg" (Introduction)<br>Beiträge teilweise französisch, teilweise englisch |
| ISBN: | 978-3-631-81137-5<br>Weitere Ausgaben: 978-3-631-83632-3, 978-3-631-83633-0, 978-3-631-83634-7 |
| Identifier: | DOI: 10.3726/b17628 |
| DOI in Druckwerken: | 10.3726/b17628 |
| Schlagwörter: | *Dichtersprache / Phraseologie / Literarischer Stil |
| Klassifikation: | Dewey Decimal Classification: 808 |
| Links zum Titel: | Inhaltstext |

WorldCat

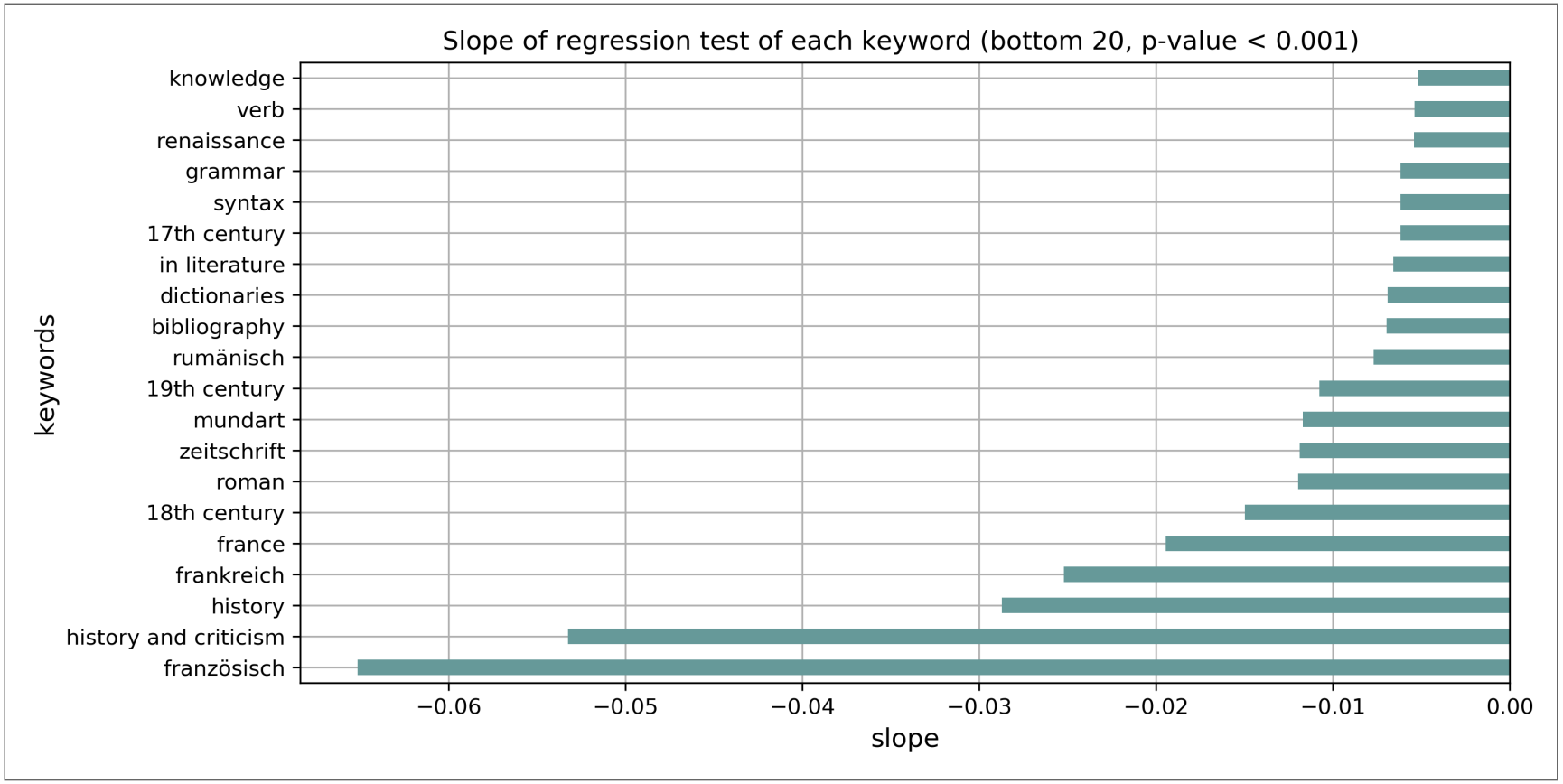
- Intellektuelle Annotation durch Fachleute

# Schlagwörter: slope (top)




- Spanisch, Italienisch
- Rezeption, Film, E-Books, Korpus
- Sprachwandel, Sprachkontakt, Mehrsprachigkeit

# Schlagwörter: slope (bottom)




- Auf und über Französisch und Rumänisch
- Klassische Gattungen und Medien
- 15. bis 19. Jahrhundert

# Schlussfolgerungen

## Schlussfolgerungen: Tendenzen

- Entwicklung der Themen
- Längere Publikationen
- Weniger Veröffentlichungen in französischer Sprache
- Mehr auf **Deutsch**, Spanisch und Englisch
- Ort der Veröffentlichung im deutschsprachigen Raum
- Gedrucktes immer noch sehr dominant
- E-Books nehmen langsam zu
- E-Books sind deutlich teurer

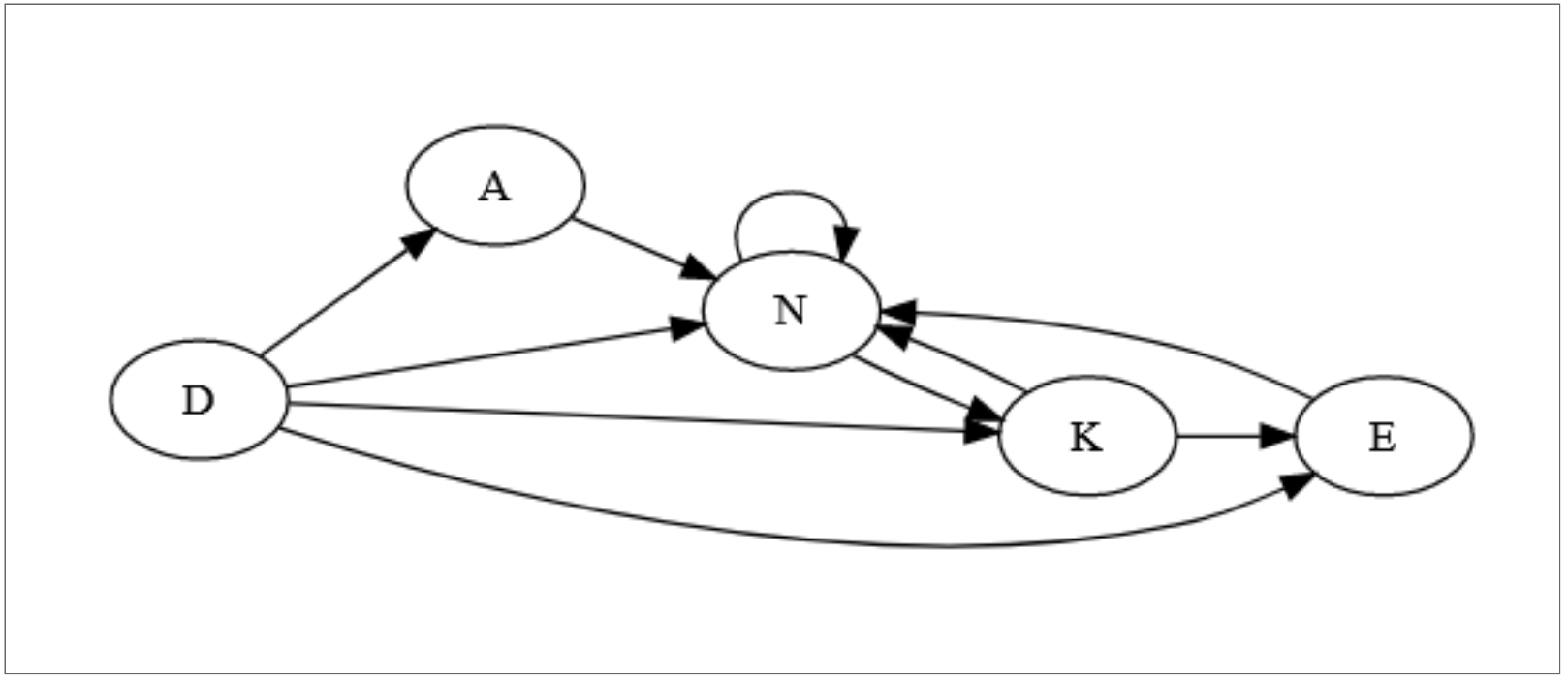
## Schlussfolgerungen: Datensatz

- Enormes Potenzial von Daten aus Katalogen
- Kombination
  - Nutzung von kuratierten Daten
  - Anwendung von quantitativen Methoden
- Mit realistischen Erwartungen
- Fachkenntnisse von mehreren Bereichen


Library
Qualitative Data
Quantitative Approaches
imgflip.com

## Danke an alle, die die Daten im Katalog

- eintragen
- bearbeiten
- annotieren
- überprüfen
- (sach)erschließen
- verwalten
- konvertieren
- ...


A
N
D
K
E