

# GeMTeX

## *German Medical Text Corpus*

Methodenplattform

Martin Boeker



2023-10-10

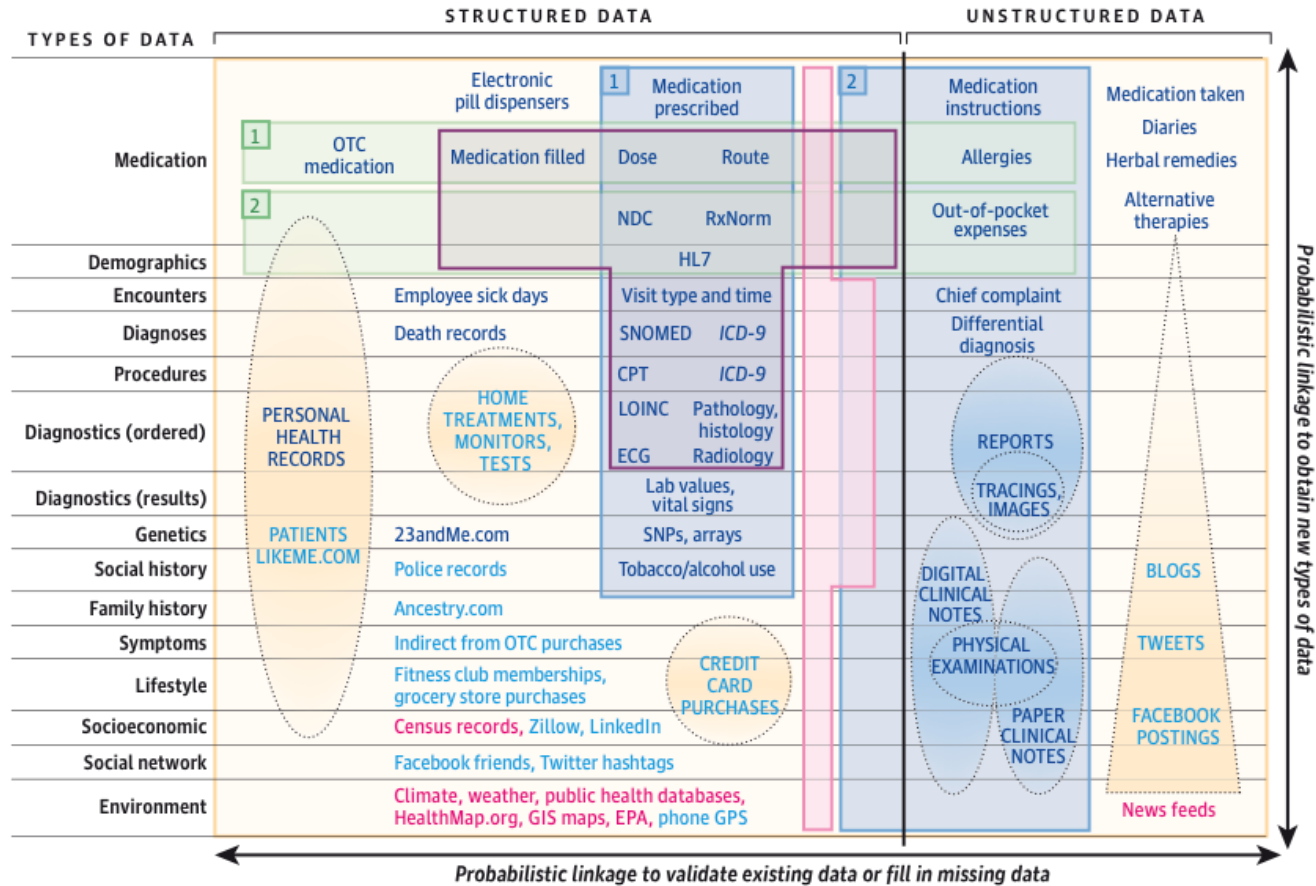
MIRACUM-DIFUTURE Symposium Erlangen



GEFÖRDERT VOM



# Strukturierte vs. unstrukturierte Daten - hoher Anteil textbasierter Daten

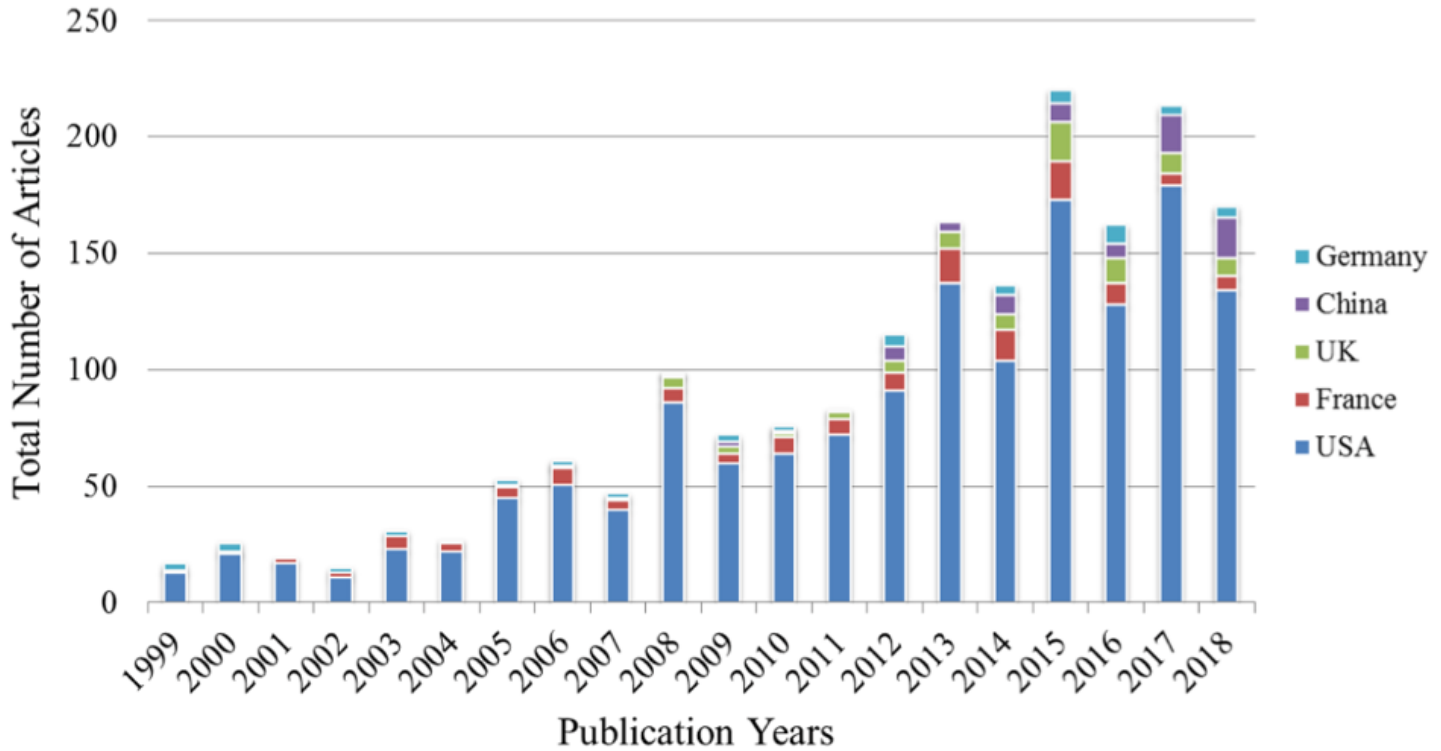


# Charakteristika klinischer Texte

Phänomen	Beispiel	Erläuterung
Telegrammstil	“Weitere Abklärung auf Intensiv”	Unvollständige Sätze, skizzenhafte, stichwortartige Ausdrucksweise
Umgangssprachlichkeit	“Coronaverdacht”, “Leberlatte”	Oft abhängig vom klinischen “Milieu”
Ad-hoc-Abkürzungen	“lymphozyteninfiltr.”	Weglassen des Wortendes mit oder ohne Punkt
Mehrdeutige Akronyme (In Kliniktexten selten definiert.)	“LCS”	“Long-Covid-Syndrom”, aber auch Bezeichnung einer Knieprothese (“low contact stress”) oder “Liquor cerebrospinalis”.
Kurzformen lokaler oder regionaler Bedeutung	“UKE” “St. n.” “EBA”	“Universitätsklinikum Hamburg-Eppendorf”; “Status nach” = “Zustand nach” (Schweiz) Interdisziplinäre Notfallambulanz in Graz (“Erstuntersuchung-Beobachtung-Aufnahme”)
Konventionalisierte Kurzformen durch externe Vorgaben	“V mors can dig V dext”	“Vulnus morsum canis digiti quinti dextri” = “Hundebisswunde am rechten Kleinfinger” (Nomenklatur der österr. Unvallversicherung)
Schreib- und Tippfehler	“Astra-Seneca-Impfung”, “Schüsselbein”, “Colonkrzinom”	Akzidentell oder systematisch (z.B. durch Nicht-Muttersprachler)
Schreibvarianten	“cervikal”, “Oesophagus”	Eindeutschungsregeln werden durch nicht oder nur teilweise beachtet
Nominalkomposita	“Außenmeniscusscheibendeformität” “Ibuprofenintoxikation”	lexikalisch nicht erfasste Langwörter durch Zusammenschreibung
Anaphern	(i) “Adeno-Ca Rectum pN+MX G2 (...). Tumor in toto exzid.” (ii) “Im Magen kein Blut (...). Zahlr.Schleimhauterosionen”	Präzise Bedeutung erschließt sich nur durch Bezugnahme auf den Vortext, in (i) meint “Tumor” das zuvor exakt beschriebene Karzinom; in (ii) sind “Schleimhauterosionen” “Erosionen der Magenschleimhaut”.
Negationen	“Kein Anhalt für Pneumonie”, “Pulmones: nihil” “metastasenfrei”	Oft jargontypische Wendungen
Unsicherheit, Verdacht	“V.a. Myokardinfarkt DD Lungenembolie”	Durch Kürzel wie “V.a.” (“Verdacht auf”) oder DD (“Differentialdiagnose”) ausgedrückt
Zeitbezüge	“Z.n. Covid-19”, “Streptokokkenangina 06/16”	“Z.n.” (Zustand nach) verweist auf früheres Ereignis, häufig Zeitangaben im Format “MM/YY”
Sonstige Kontexte	(i) “Vater: Pankreas-Ca” (ii) “Von Bauchlagerung wurde Abstand genommen”	(i) Familienanamnese bezieht sich auf Erkrankungen von Angehörigen. (ii) nicht ausgeführte Planungen

# Publikationen: NLP in der Medizin, Nationalität Erstautor

n = 2336, aus Deutschland 2,44 %



Wang, J. et al. Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed. J Med Internet Res 22, e16816 (2020).

## Wichtigste Ursachen

1. Fehlenden **verfügbare** deutschsprachigen klinische Korpora
  - a. Rechtliche Grundlagen
2. Wenige übersetzte Vokabularien/Terminologien
  - a. aber: Interfaceterminologien verfügbar
3. Werkzeuge nicht optimal an deutsche Sprache angepasst

CAVE: Forschungsinteressen spiegeln nicht unbedingt operational eingesetzte Systeme

Corpus	Text Genre	# Documents	# Sentences	# Tokens	Shareability
3000PAJ [6]	Discharge summaries	1,106	146,191	1,707,019	Non-Shareable
JSYNCC OP [5]	Medical text-books	399	20,860	199,569	Code for re-creation
GGPONC 1.0 [3]	Clinical practice guidelines	12,761	77,986	1.522,588	DUA
BRONCO150 [4]	Discharge summaries	150	10,251	83,633	DUA
<b>This work</b>	<b>Alienated case reports</b>	<b>63</b>	<b>5,430</b>	<b>43.667</b>	<b>Fully Shareable</b>

Hahn, U., Matthies, F., Lohr, C., Löffler, M. 3000PA—Towards a National Reference Corpus of German Clinical Language. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth 26–30 (2018) doi:10.3233/978-1-61499-852-5-26.

Lohr, C., Buechel, S. & Hahn, U. Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus. in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (European Language Resources Association (ELRA), 2018).

Borchert, F. et al. GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines. (2020).

Kittner, M. et al. Annotation and initial evaluation of a large annotated German oncological corpus. JAMIA Open 4, (2021).

Luise Modersohn, Stefan Schulz, Christine Lohr, & Udo Hahn. GraSCCo — The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. in GMDS 2022 (2022).

## scientific **data**

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [scientific data](#) > [data descriptors](#) > [article](#)

Data Descriptor | [Open Access](#) | [Published: 14 April 2023](#)

### **A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters**

[Phillip Richter-Pechanski](#) , [Philipp Wiesenbach](#), [Dominic M. Schwab](#), [Christina Kiriakou](#), [Mingyang He](#), [Michael M. Allers](#), [Anna S. Tiefenbacher](#), [Nicola Kunz](#), [Anna Martynova](#), [Noemie Spiller](#), [Julian Mierisch](#), [Florian Borchert](#), [Charlotte Schwind](#), [Norbert Frey](#), [Christoph Dieterich](#) & [Nicolas A. Geis](#)

[Scientific Data](#) **10**, Article number: 207 (2023) | [Cite this article](#)

**361** Accesses | **1** Altmetric | [Metrics](#)

### **Abstract**

We present CARDIO:DE, the first freely available and distributable large German clinical corpus from the cardiovascular domain. CARDIO:DE encompasses 500 clinical routine

# Hintergrund für GeMTeX

- Vielzahl nicht strukturiert erschlossener Informationen in *klinischen* Texten
  - Arztbriefe
  - Befundberichte (z. B. Bildgebung, Pathologie)
  - Berichte über Prozeduren (z. B. OP)
  - Anamnesen
- Erfolgreiche Nutzung von NLP besonders im englischsprachigen Raum
- NLP an deutschsprachigen klinischen Texten: vergleichsweise geringe Fortschritte
- Größtes Hindernis in Deutschland und für die MII:  
*Fehlen von deutschsprachigen (annotierten) klinischen Texten (Textkorpus)*



# Zielsetzung GeMTeX

- Deutsches medizinisches (klinisches) Referenz-Korpus der MII
- (Prospektive) Textdaten als Ressource für die Forschung
  - Semantische Goldstandard Annotationen
  - Trainierte Sprachmodelle
  - Algorithmische Auswertung
- Nutzung von NLP im Rahmen der DIZ
- Initialisierung von Folgevorhaben
  - Demonstration von Vorteilen der semantischen Textanalyse für die Krankenversorgung

# Anknüpfungspunkte GeMTeX an MII und NUM

Methodenplattform mit

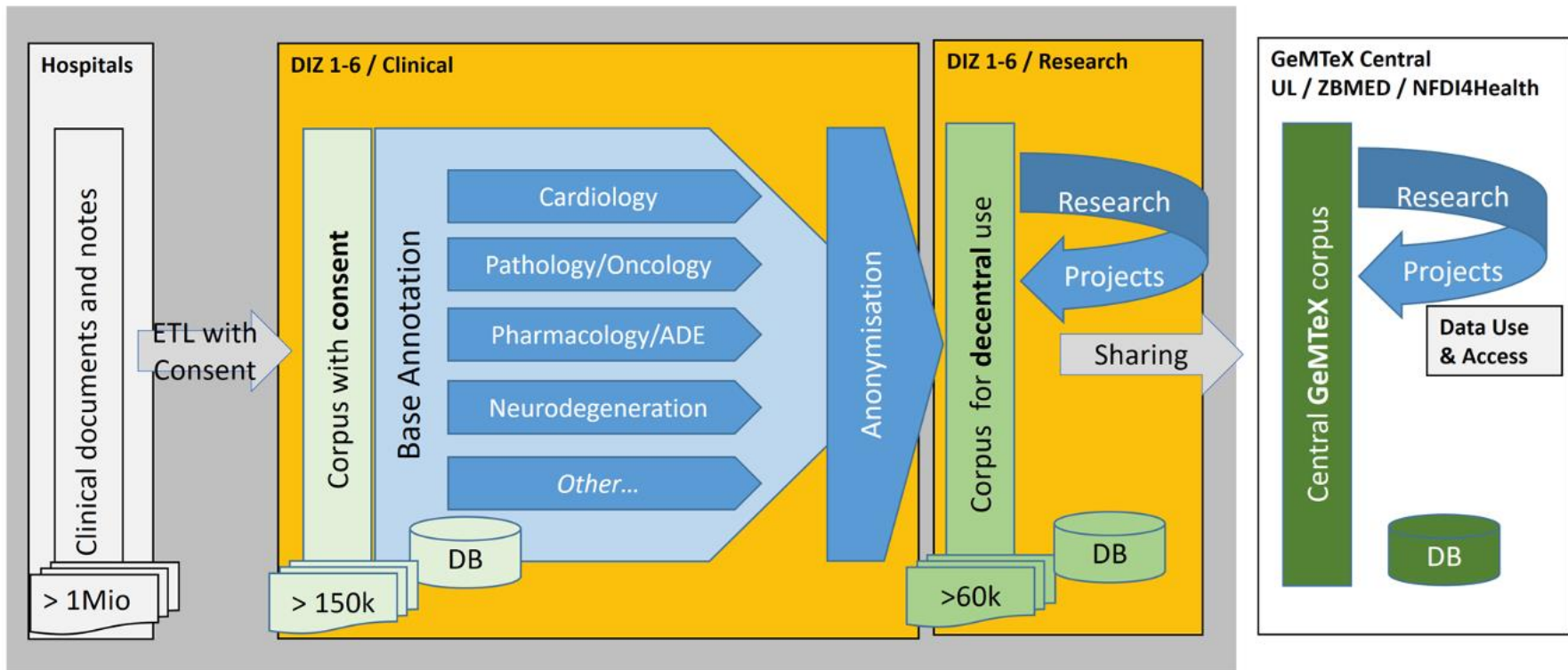
- Erweiterung des Datenschutzkonzeptes für (de-identifizierte) Texte
- Nutzung des Broad Consent und Erweiterung der Governance
- Standardisierung der Nutzung von Texten
- Definition unterschiedlicher Integrationskonzepte (Nutzungsszenarien)
- Bereitstellung von Methoden
  - Nutzung von Texten
  - Annotation
  - Training von Modellen
- Anwendungen von (förderierten) Verfahren der KI

# GeMTeX Partner

- Technische Universität München
- Universität Leipzig/ Universitätsklinikum Leipzig
- TU Darmstadt
- Universitätsmedizin Essen
- Charité Berlin
- Universitätsklinikum Erlangen
- Universitätsklinikum Dresden
- Universitätsklinikum Heidelberg
- Universität Münster
- Hasso-Plattner-Institut
- Medizinische Hochschule Hannover
- Ludwig-Maximilians-Universität München
- Informationszentrum Lebenswissenschaften ZB MED
- Universitätsklinikum Tübingen
- Averbis GmbH
- ID Berlin
- Medizinische Universität Graz
- Friedrich-Schiller-Universität Jena

- 16 Partner
  - Integration der NWG NLP DE.xt
- 2 assoziierte Partner
- Förderung 6.8 Mio. €
- Laufzeit 3.5 (3) Jahre





Annotation Project (DIC specific)

Annotation Database

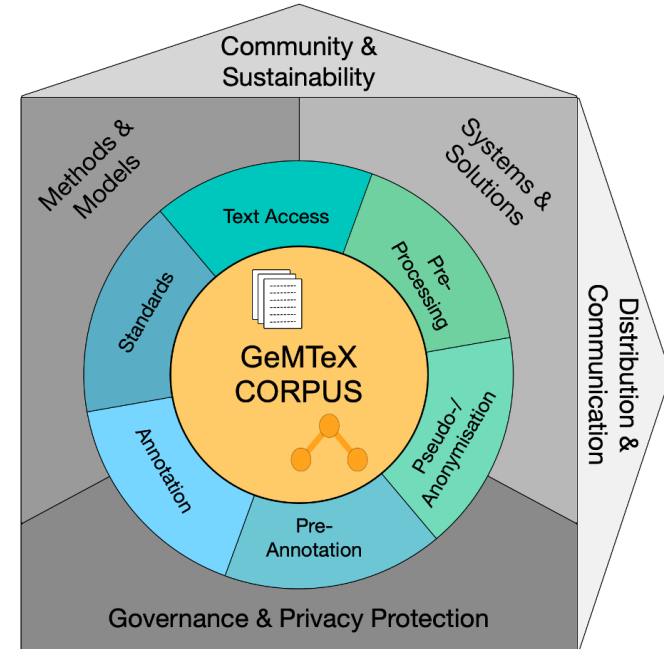


summarized # of texts over all 6 sites

DIC/Site area

# Aufbau eines (deutschsprachigen) klinischen Korpus

- Ausleitung von Textdokumenten und textuellen Inhalten aus den KIS
  - in allen Formen
  - Tooling wird zur Verfügung gestellt
- De-Identifikation/Anonymisierung
- Annotation
- Nutzungs-Integration
  - lokal
  - zentral aggregiert
  - föderiertes Lernen und Modellintegration
- Governance und Privatheit
- Evaluation des Korpus
- Standardisierung der Prozesse und Repräsentation





# Systemaufbau an Annotations-Standorten

- Bereitstellung Averbis Health Discovery
- Bereitstellung INCEpTION Annotationsplattform
- Bereitstellung sichere Arbeitsumgebungen für Annotatoren
  - Möglichst mit remote Zugriff über VPN

# Averbis Health Discovery

- Steht allen Partnern in GeMTeX zu Verfügung
- Viele Werkzeuge vorhanden
  - De-Identifikation
  - Annotation
    - Diagnosen
    - Medikation
- Workflow-Engine enthalten
  - Pipelining von Prozessschritten
  - Python-API ermöglicht Einbinden eigenen Tools
- Averbis unterstützt die Standorte







# Annotationseditor: INCEpTION

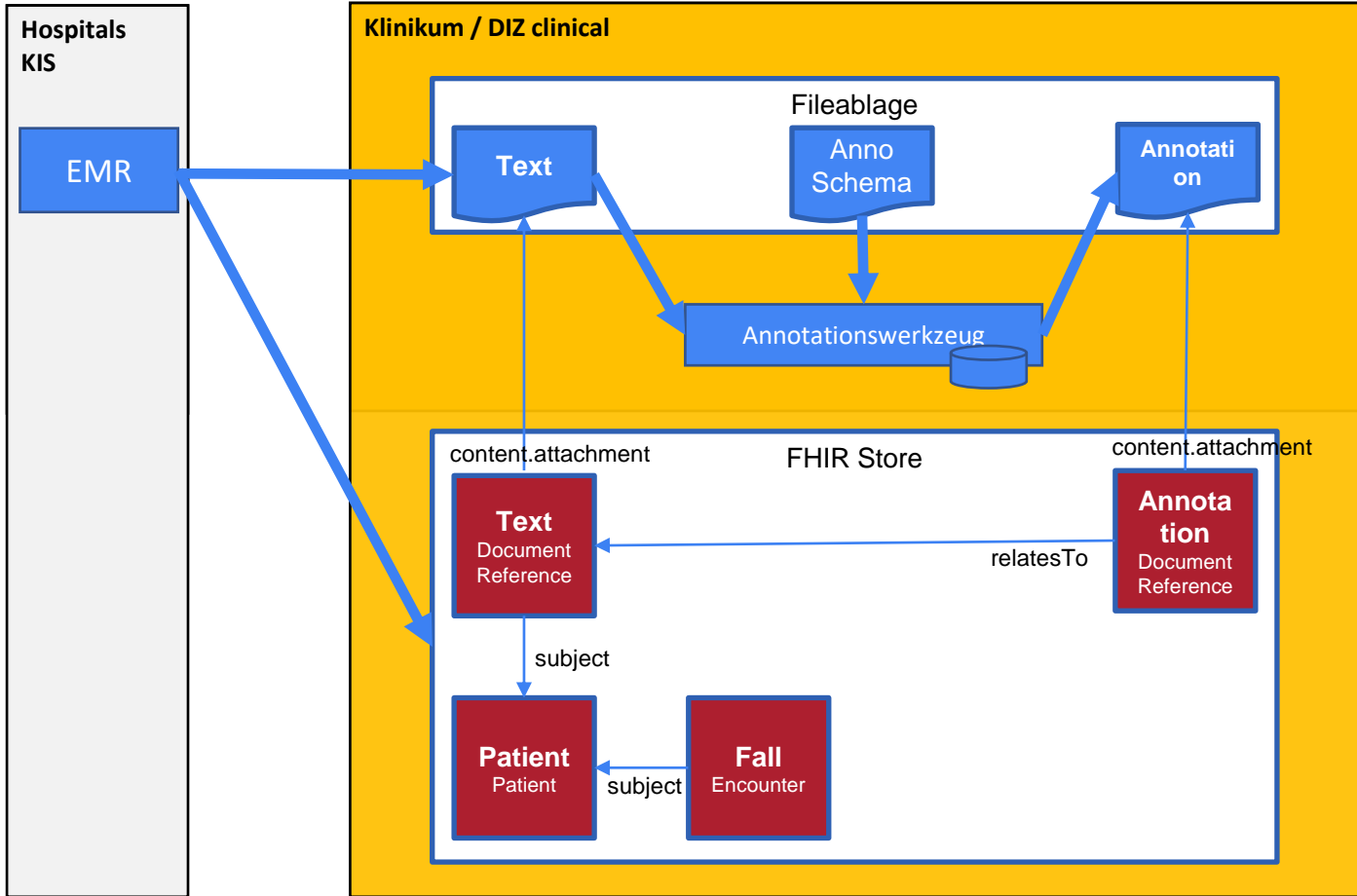
- Kooperation mit TU Darmstadt – Richard Eckart de Castilho & Iryna Gurevych
- <https://inception-project.github.io/>
- Anpassungen für GeMTeX geplant
  - Projektmonitoring der Annotation
  - Einbindung von Annotationsvokabularien
  - Einbindung Prä-Annotation mit interaktivem Lernen

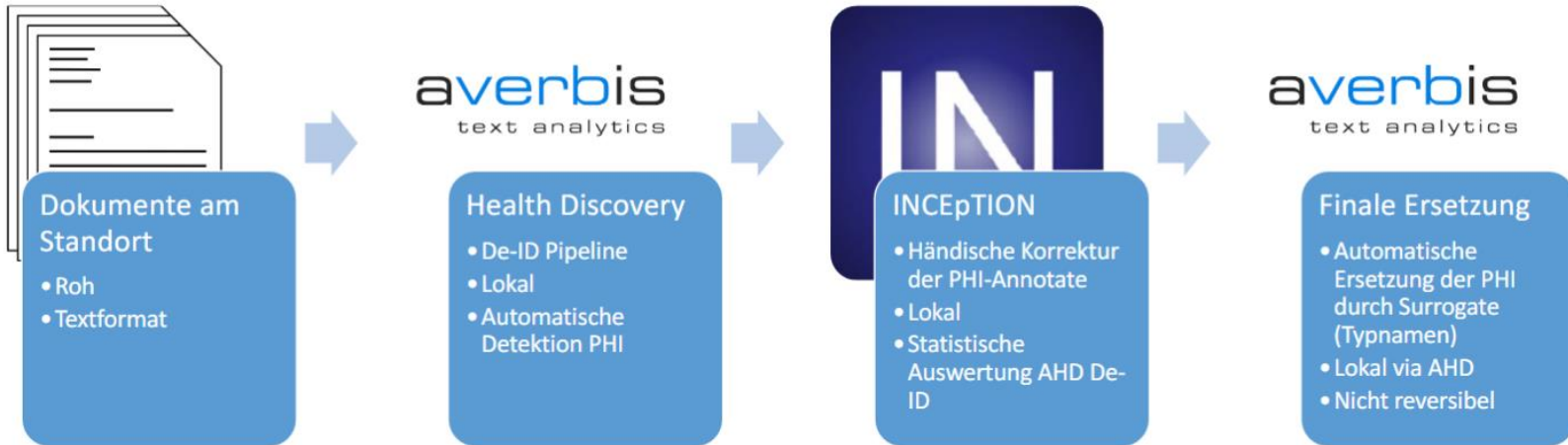
The screenshot displays the INCEpTION annotation editor interface. The main window shows a text document with various annotations. The text includes: "Barack Obama II PERI", "born August 4, 1961)", "is an", "American politician who served as the 44th", "President of the United States from 2009 to 2017", "The first African American to assume the presidency, he was", "previously the junior United States Senator from Illinois from 2005 to 2008.", and "He served in the Illinois State Senate from 1997 until 2004.". Annotations include "Barack Obama II PERI", "born August 4, 1961)", "is an", "American politician who served as the 44th", "President of the United States from 2009 to 2017", "The first African American to assume the presidency, he was", "previously the junior United States Senator from Illinois from 2005 to 2008.", and "He served in the Illinois State Senate from 1997 until 2004.". A sidebar on the left shows the "Active Learning" section with a "Layer" dropdown set to "Named entity", a "Recommendation" section with "Text" set to "Illinois", and a "Learning History" table. A sidebar on the right shows the "Annotation" section with a "Layer" dropdown set to "Named entity", a "Text" field set to "Illinois", and an "Identifier" field set to "illi". A dropdown menu for "Identifier" is open, showing options: "Illinois Senate", "Illinois River", "Governor of Illinois", "Alton", "Illinois Country", and "Illinois Territory". A tooltip for "Illinois Senate" is visible, showing its definition: "upper chamber of the Illinois General Assembly, the legislative branch of the government of the state of Illinois in the United States".

Learning History	Layer	Label	Score	Delta	Action
Berkeley <a href="http://www.wikidata.org/entity/Q168756">http://www.wikidata.org/entity/Q168756</a>	skipped				skipped
Tesla	PER		accepted		accepted
Tesla	PER		accepted		accepted
Tesla	PER		accepted		accepted
Tesla	PER		accepted		accepted
Tesla	PER		accepted		accepted
Science	OTH		rejected		rejected
Tesla	PER		accepted		accepted

# Dateihandling/Pseudonymisierung

- Ausleitung von Texten aus Primärsystemen
- Ablage im Dateisystem
- Normalisierungsschritte
- Standardisierung von Metadaten/Texten (FHIR)
- Pseudonymisierung
  - Generierung eindeutige Bezeichnung
  - De-Identifikation der Texte
- Monitoring dezentral/zentral





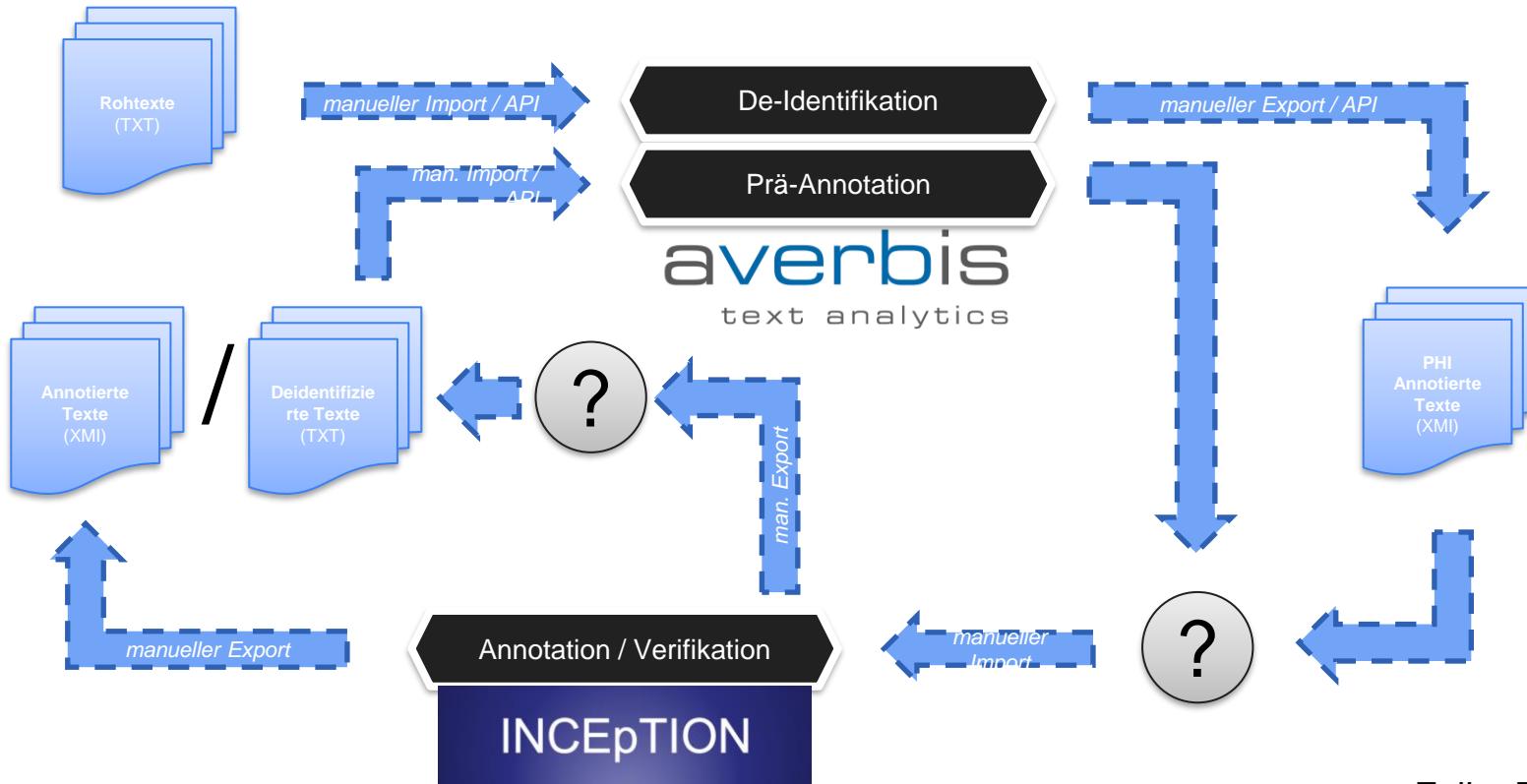
- Weitere Annotationen finden **ausschließlich** auf pseudonymisierten Texten statt
- Identifizierung der Personen **ausschließlich** über Dateinamen oder Metainformationen und Treuhandstelle
- Offset-Erhaltung **nicht** notwendig
- Ersetzung der markierten PHI-Elemente durch Typonamen (evtl. inklusive Länge des Ursprungstoken) via AHD

# Annotationsmethodik



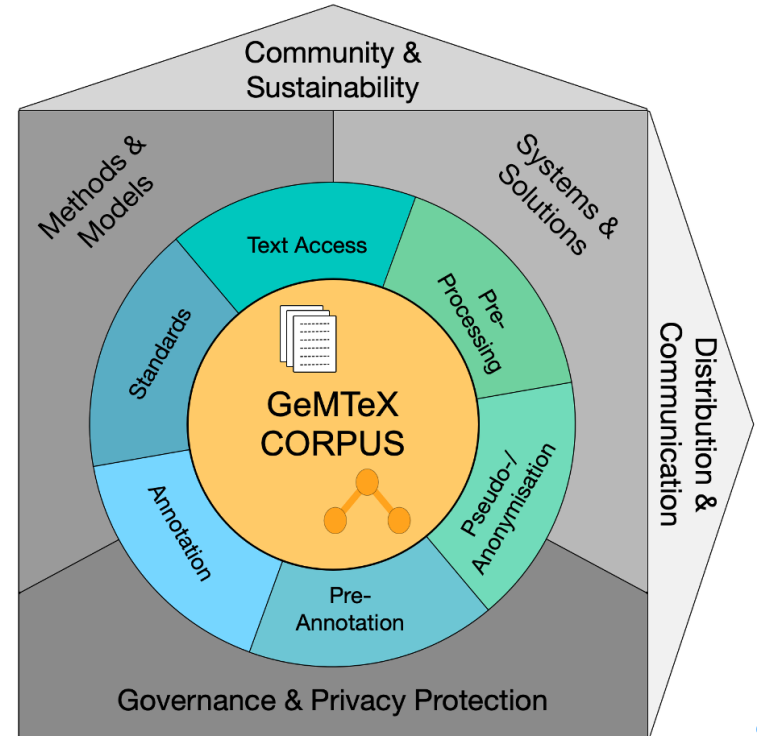
- Aufbauend auf bestehenden Methoden/Werkzeugen
  - Annotationsguideline(s) – basierend auf bestehenden GL (AIDAVA, international) & Vorarbeiten Jena
  - Annotationsterminologie(n) – SNOMED CT und LOINC
  - Automatisierte Vorannotation – AHD und spez. Projekte (HD, HPI)
  - Annotationseditor (INCEpTION)
- Annotation mit Medizinstudenten an den Standorten
  - Mindestens Teams von 10 Studierenden erforderlich
    - Frühzeitiger Aufbau der Teams
  - Dokumentar:innen zur Ltg. der Teams/Qualitätsüberwachung
- Integration von interaktivem Lernen
- Annotation
- Qualitätsprüfung Annotation und De-Identifikation

# IMISE Referenzplattform



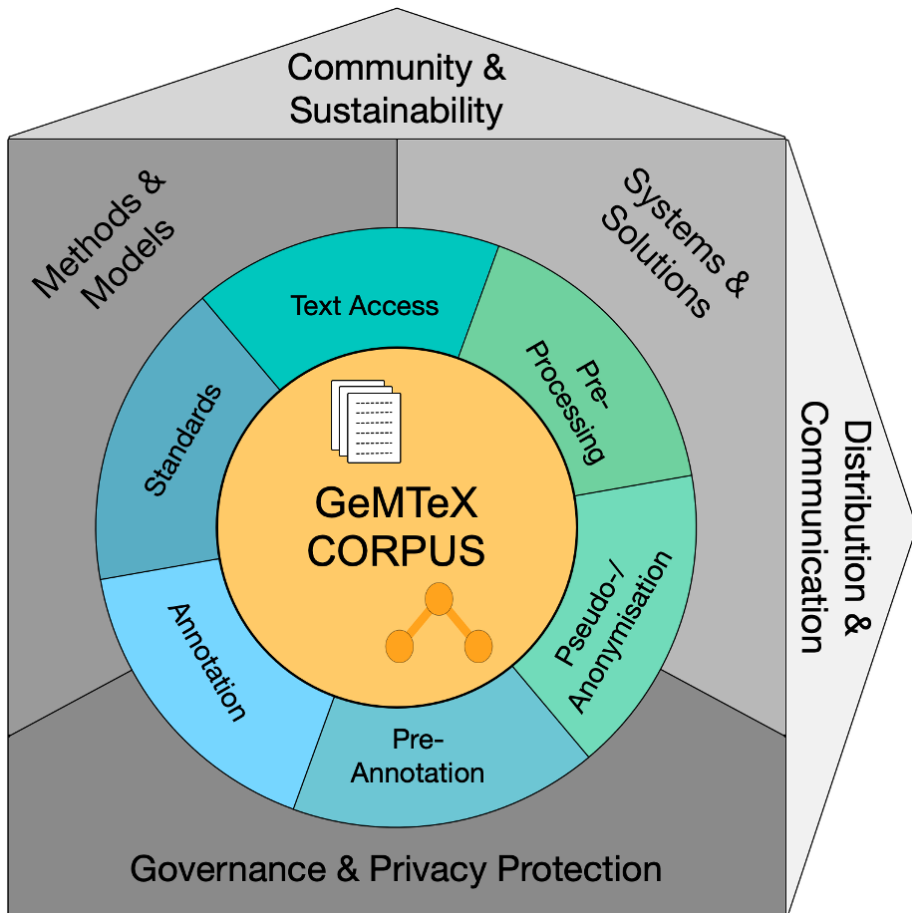
# GeMTeX - Zusammenfassung

- Bereitstellung eines deutschen klinischen Text-Corpus
  - Umfangreich (60 k annotierte Dokumente)
  - Standardisiert
  - Definierte Qualität
  - Offen
- Governance der MII
- 16 Partner
- Nachhaltigkeit durch Community-Integration
- Nutzung von LLM?







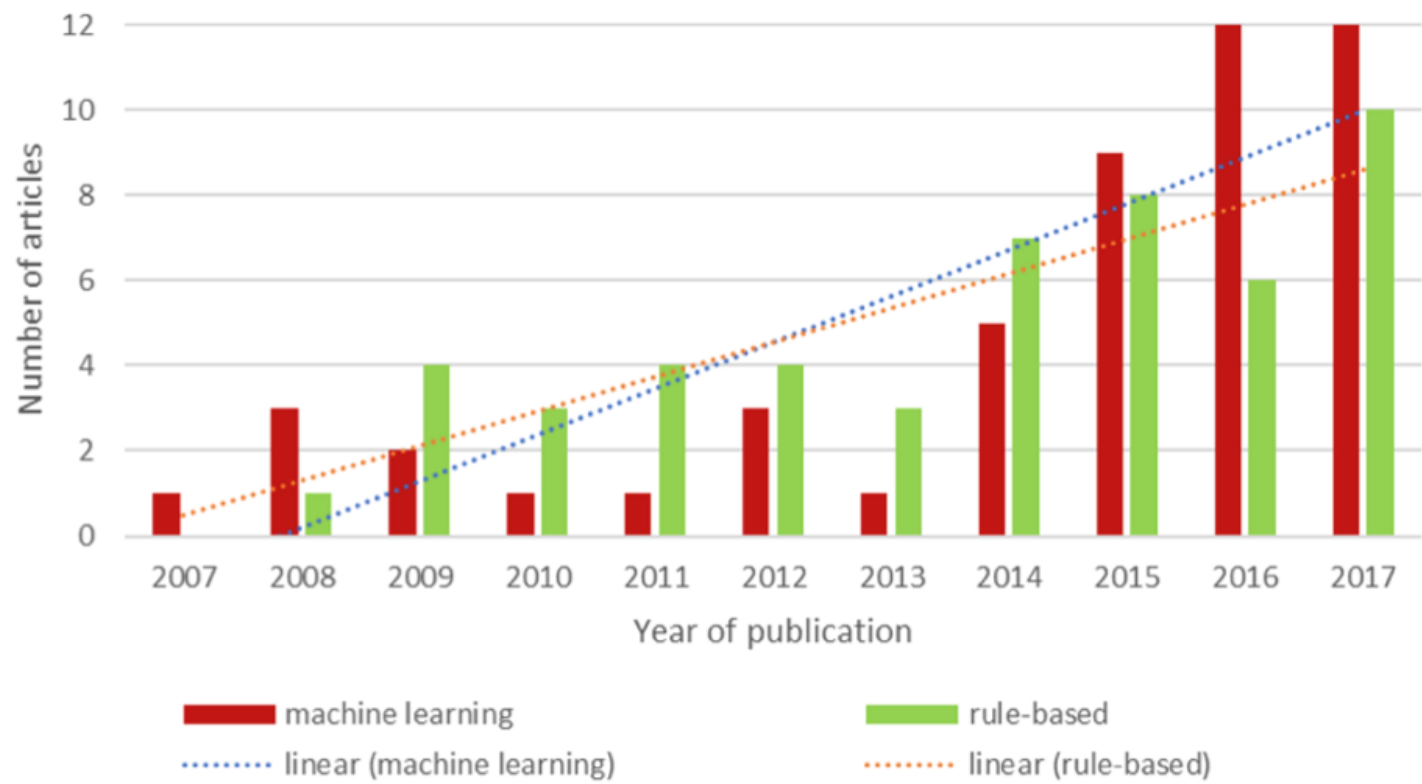


Corpus	Documents	Sentences	Types	Tokens	Available
Wermter and Hahn (2004) (FRAMED)	–	6,494	20,729	100,150	✗
Fette et al. (2012)	544	–	–	–	✗
Bretschneider et al. (2013b)	174	4,295	3,979	28,009	✗
Bretschneider et al. (2013a)					
Toepfer et al. (2015)	140	–	–	–	✗
Lohr and Herms (2016)	450	22,427	11,008	266,390	✗
Kreuzthaler and Schulz (2015)	1,696	–	–	–	✗
Kreuzthaler et al. (2016)					
Roller et al. (2016)	1,725	<b>27,939</b>	–	158,171	✗
Cotik et al. (2016)	183	2,234	–	12,895	✗
Krebs et al. (2017)	<b>3,000</b>	–	–	–	✗
Hahn et al. (2018) (3000PA)	<b>3,000</b>	–	–	–	✗
<b>JSYNCC (this work)</b>	867	24,895	<b>32,108</b>	<b>312,784</b>	✓

Table 1: Overview of existing corpora of German clinical language. Highest value per column in bold.

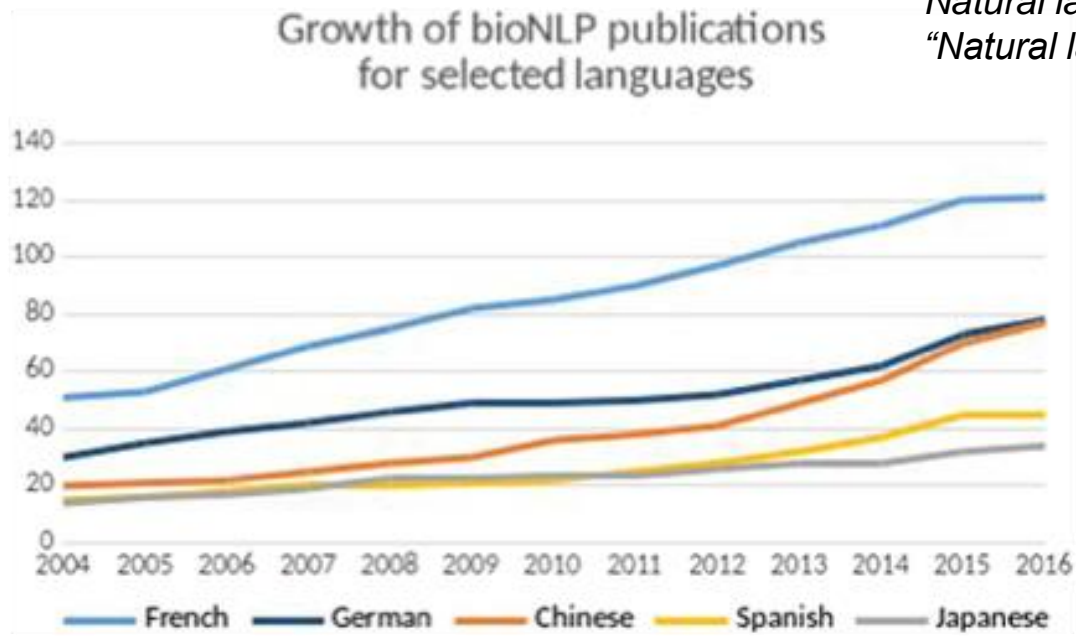
# Regelbasierte Ansätze vs. ML in NLP: Paradigmenwechsel (2019)

n = 106



# Klinisches NLP in nicht-englischen Sprachen (2018)

Natural language processing AND \*languag\*[tiab] →  
"Natural language processing"[tw] AND German\*[tiab]



n = 435

aktuell 121 German

**Fig. 1** Growth of bio-clinical NLP publications in MEDLINE over the past decade, for the top 5 studied languages other than English

Névél, A., Dalianis, H., Velupillai, S., Savova, G. & Zweigenbaum, P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semant 9, 12 (2018).