DIFUTURE
Data Integration for Future Medicine

HiGHmed
Medical Informatics

miracum
Medical Informatics in Research and Care in University Medicine

SMITH
Smart Medical Information Technology for Healthcare

GeMTeX

German Medical Text Corpus

Methodenplattform

Martin Boeker

MRI TUM

MEDIZIN INFORMATIK INITIATIVE

2023-08-15 MIRACUM-DIFUTURE Kolloquium

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Zielsetzung GeMTeX

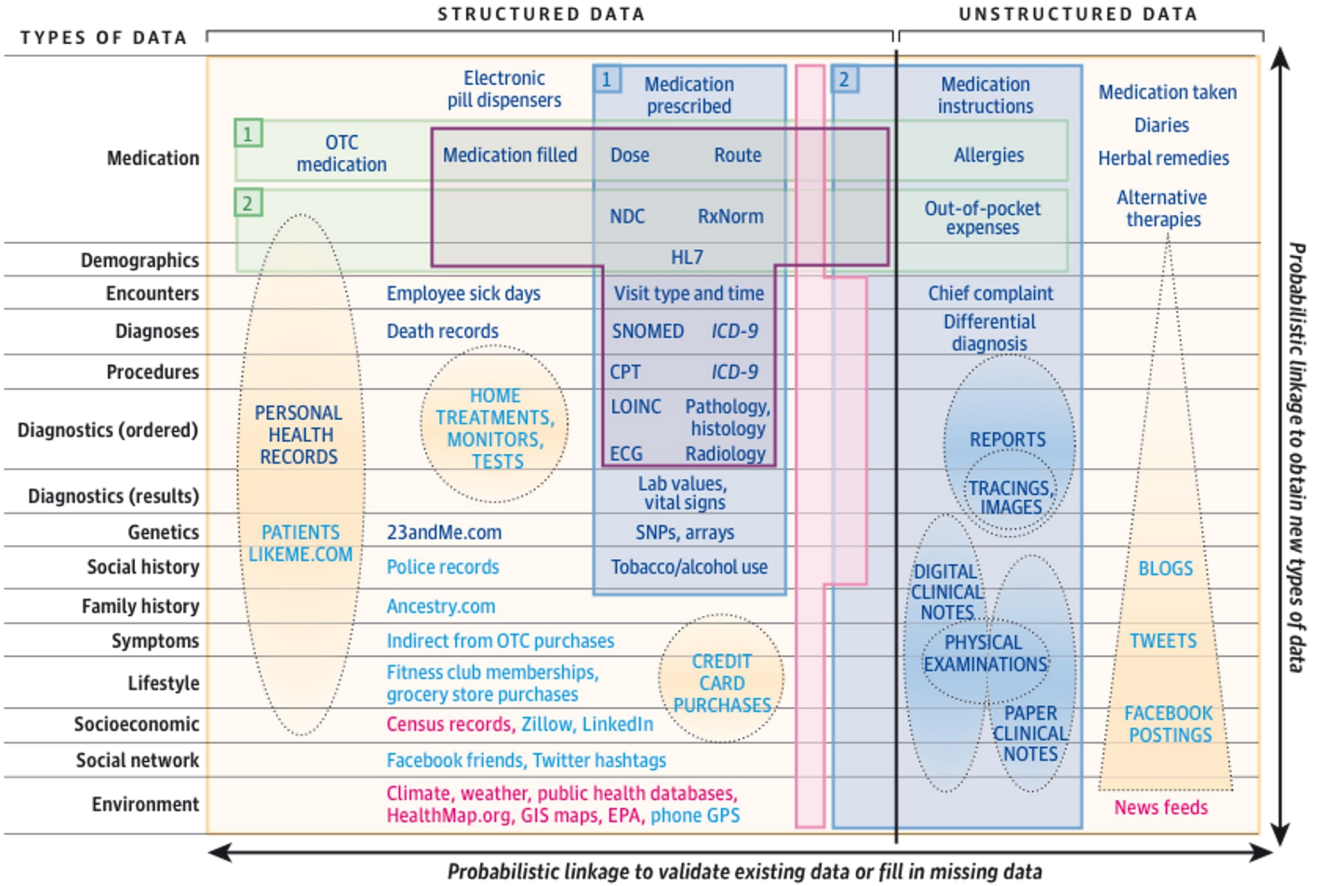
- Deutsches medizinisches (klinisches) Referenz-Korpus der MII
- (Prospektive) Textdaten als Ressource für die Forschung
  - Semantische Goldstandard Annotationen
  - Trainierte Sprachmodelle
  - Algorithmische Auswertung
- Nutzung von NLP im Rahmen der DIZ
- Initialisierung von Folgevorhaben
  - Demonstration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

Strukturierte vs. unstrukturierte Daten - hoher Anteil textbasierter Daten



Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the Missing Link for Big Biomedical Data. *JAMA* (2014) doi:10.1001/jama.2014.4228.

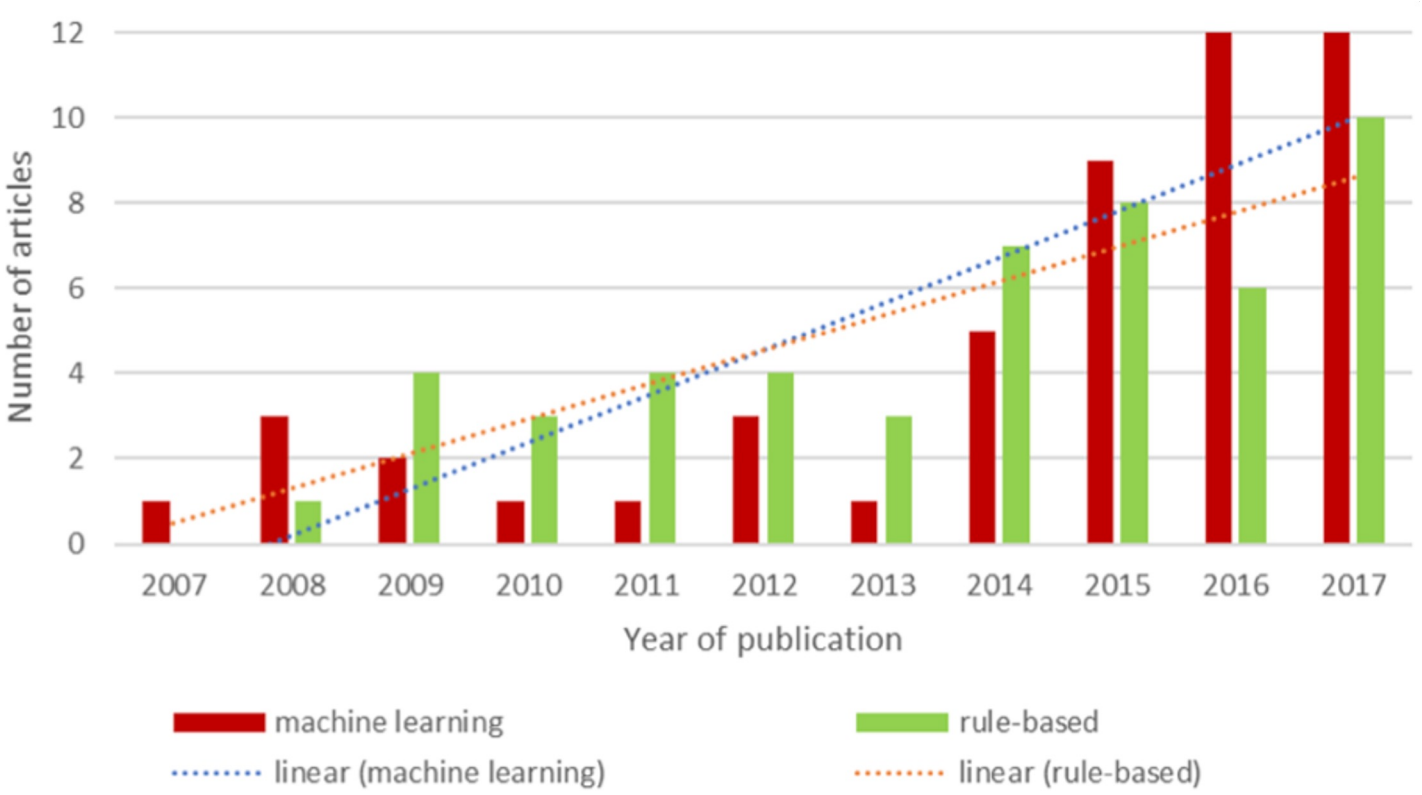
Charakteristika klinischer Texte

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

Stefan Schulz. Maschinelle Analyse klinischer Sprache in Zeiten von COVID-19. mdi 4/2021, 107–111 (2021).

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

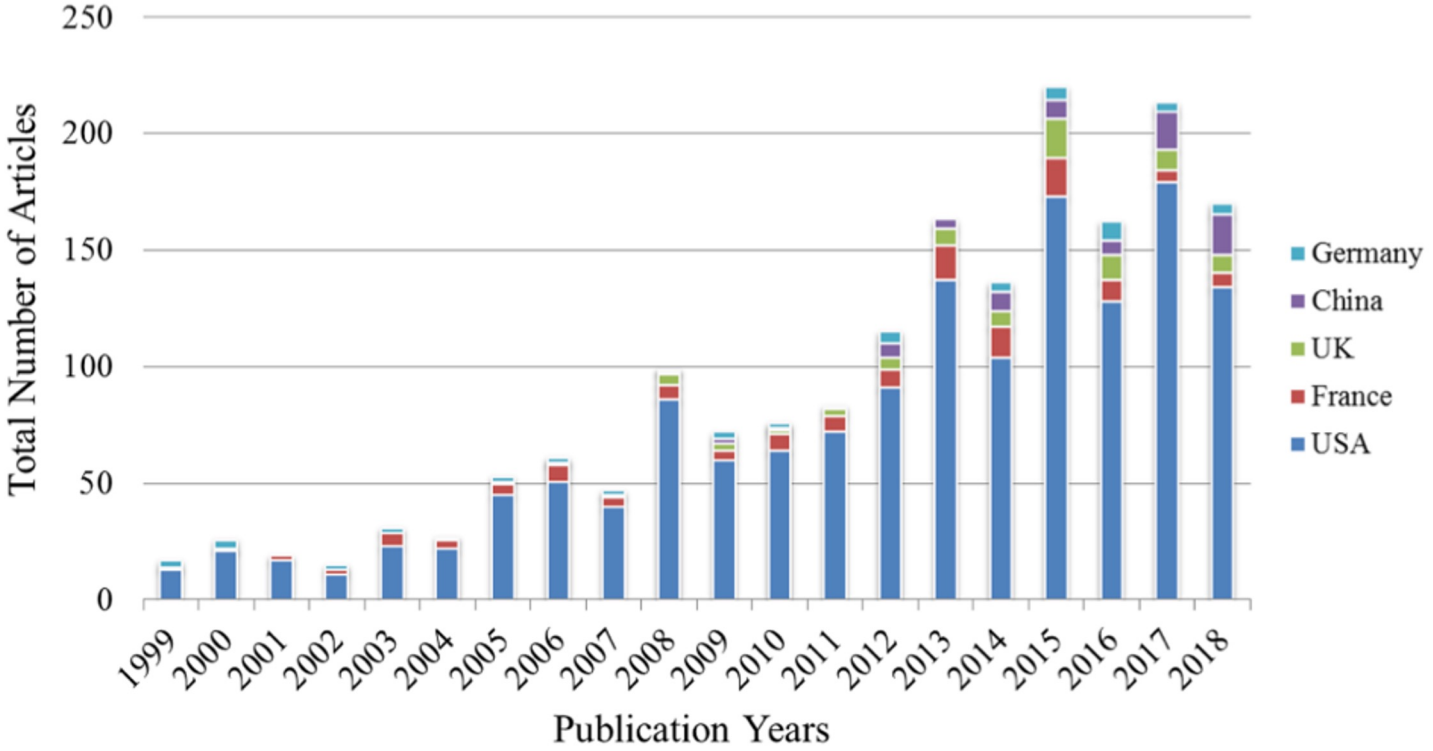
n = 106



Sheikhalishahi, S. et al. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Medical Informatics 7, e12239 (2019).

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).

Klinisches NLP in nicht-englischen Sprachen (2018)

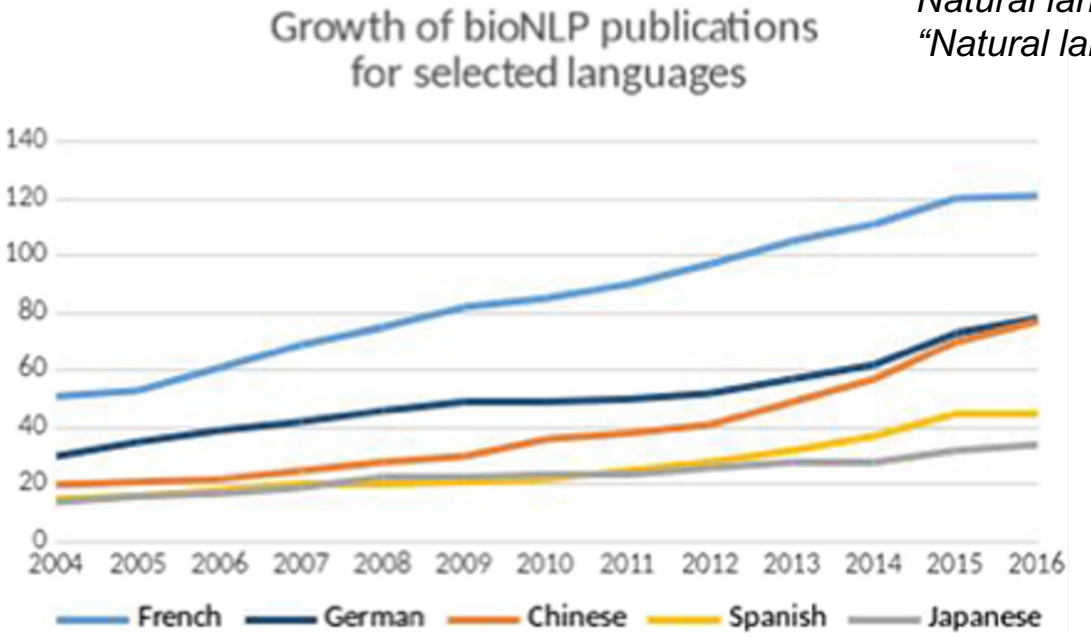


Fig. 1 Growth of bio-clinical NLP publications in MEDLINE over the past decade, for the top 5 studied languages other than English

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

n = 435

aktuell 121 German

Névéol, 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).

CORPORA bis 2018

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) Bretschneider et al. (2013a)	174	4,295	3,979	28,009	✗
Toepfer et al. (2015)	140	–	–	–	✗
Lohr and Herms (2016)	450	22,427	11,008	266,390	✗
Kreuzthaler and Schulz (2015) Kreuzthaler et al. (2016)	1,696	–	–	–	✗
Roller et al. (2016)	1,725	**27,939**	–	158,171	✗
Cotik et al. (2016)	183	2,234	–	12,895	✗
Krebs et al. (2017)	**3,000**	–	–	–	✗
Hahn et al. (2018) (3000PA)	**3,000**	–	–	–	✗
JSYNCC (this work)	867	24,895	**32,108**	**312,784**	✓

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

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).

CORPORA: Graz Synthetic Clinical Corpus- GraSCCo (2022)

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
This work	**Alienated case reports**	**63**	**5,430**	**43.667**	**Fully Shareable**

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).

Verfügbares klinisches Textcorpus – CARDIO:DE

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

Geringe Durchsetzung von NLP an deutschsprachigen klinischen Texten

Wichtigste Ursachen

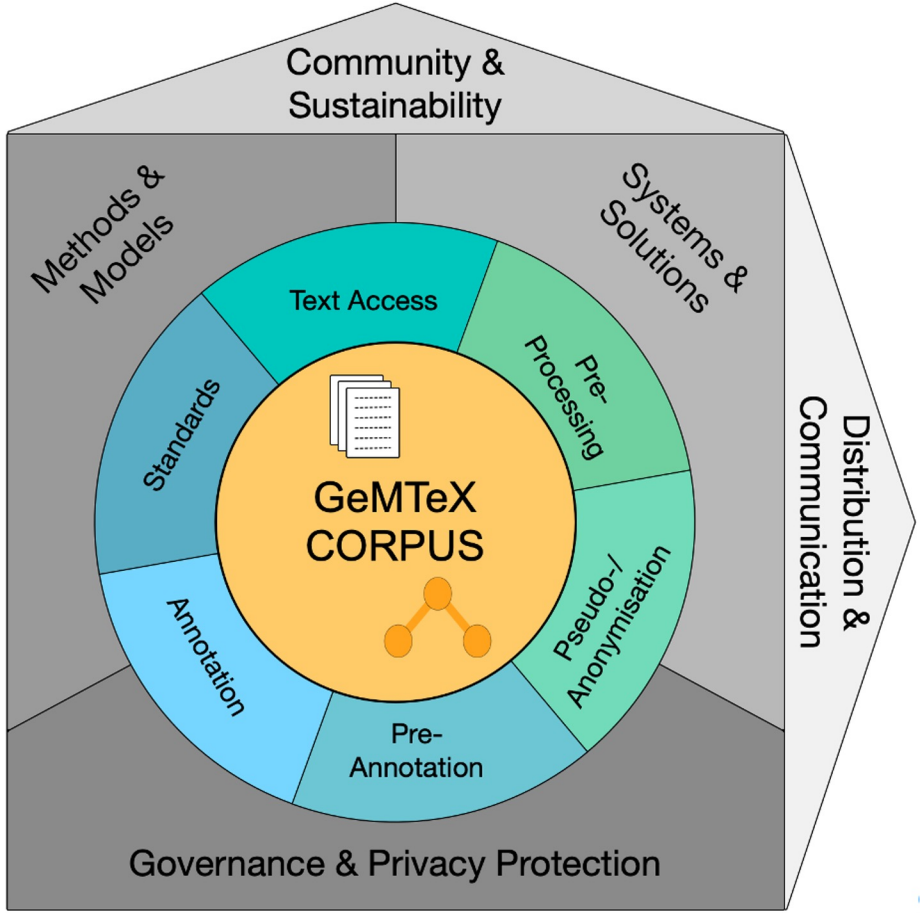
1. Fehlenden ***verfügbare*** deutschsprachigen klinische Korpora
   a. Rechtliche Grundlagen
2. Wenige übersetzte Vokabularien/Terminologien
   a. aber: Interfaceterminologien verfügbar
3. Werkzeuge nicht optimal an deutsche Sprache angepasst

CAVE: Forschungsinteressen spiegeln nicht unbedingt operational eingesetzte Systeme

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)


Community &
Sustainability
Methods &
Models
Systems &
Solutions
Distribution &
Communication
Text Access
Pre-
Processing
Standards
GeMTeX
CORPUS
Pseudo-/
Anonymisation
Annotation
Pre-
Annotation
Governance & Privacy Protection

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

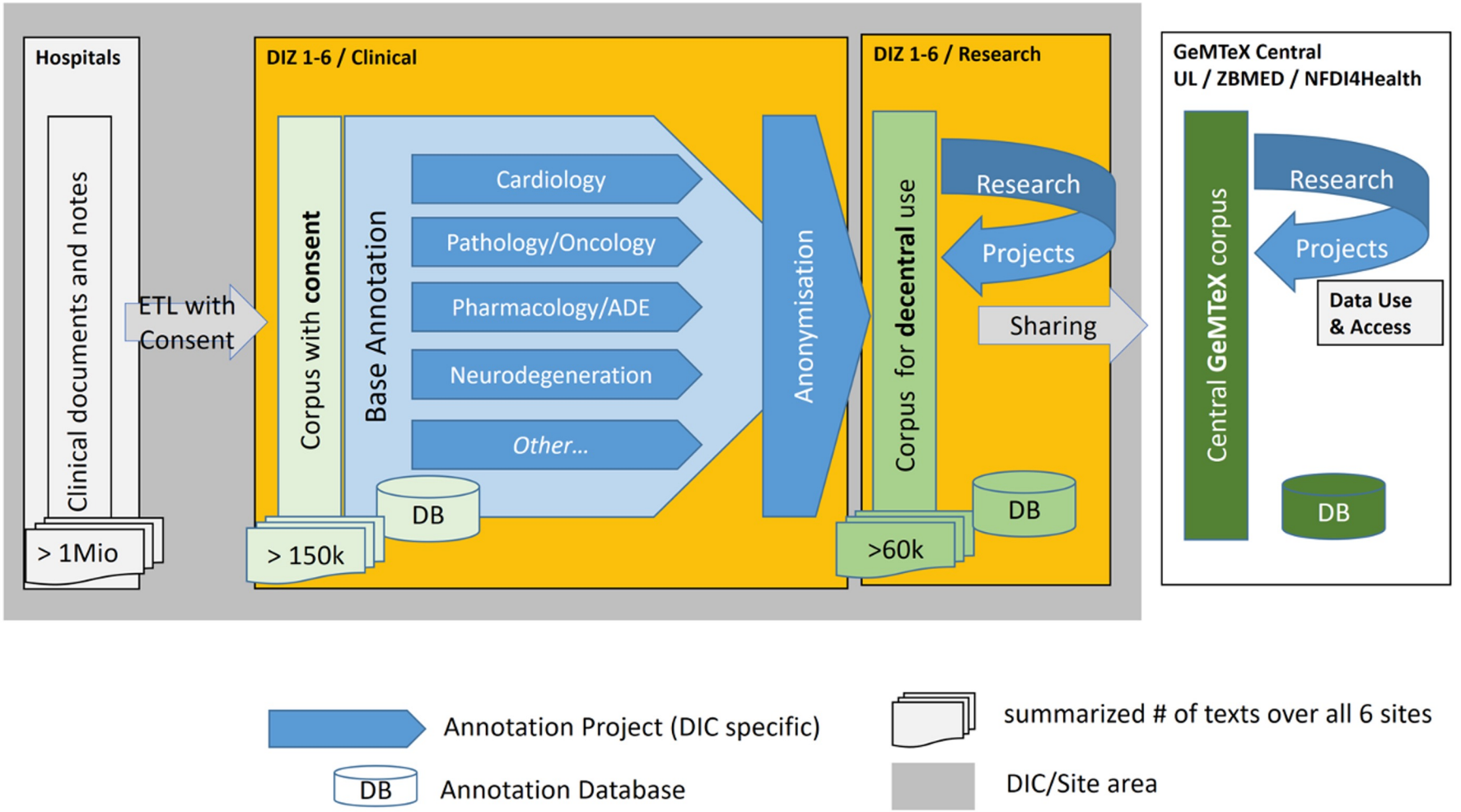
- 16 Partner
  - Integration der NWG NLP DE.xt
- 2 assoziierte Partner
- Förderung 6.8 Mio. €
- 3.5 (3) yr. Dauer

DIFUTURE
Data Integration for Future Medicine

SMITH
Smart Medical Information Technology for Healthcare

miracum
Medical Informatics in Research and Care in University Medicine

HiGHmed
Medical Informatics

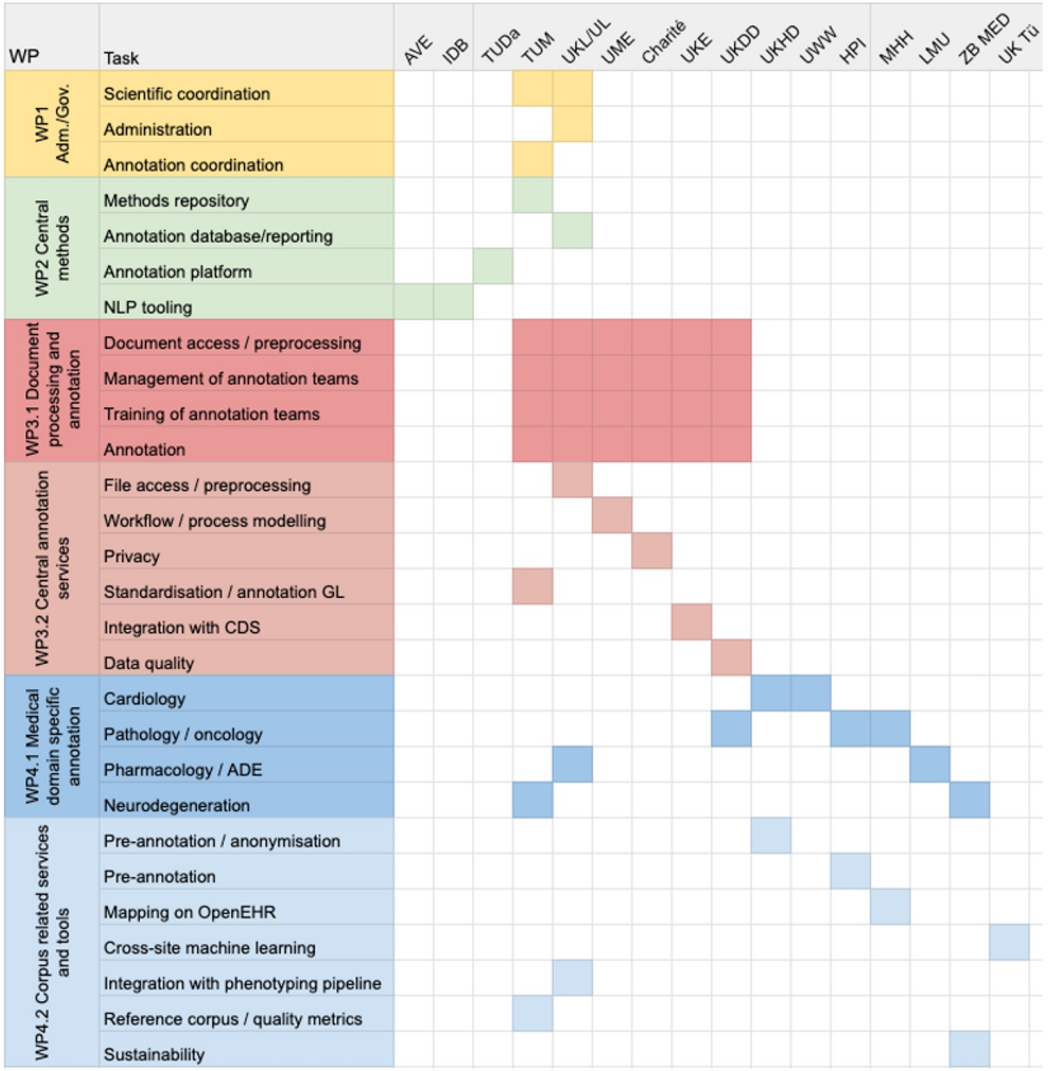

Hospitals
Clinical documents and notes
> 1Mio
ETL with Consent
DIZ 1-6 / Clinical
Corpus with consent
Base Annotation
Cardiology
Pathology/Oncology
Pharmacology/ADE
Neurodegeneration
Other…
DB
> 150k
Anonymisation
DIZ 1-6 / Research
Corpus for decentral use
Research
Projects
Sharing
DB
>60k
GeMTeX Central
UL / ZBMED / NFDI4Health
Central GeMTeX corpus
Research
Projects
Data Use & Access
DB
Annotation Project (DIC specific)
DB
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

WP-Struktur

- Komplexe Aufgaben ⇒ komplexe Struktur
  - *Annotation* und Bearbeitung von Themen zur Corpus-Nutzung an 6 Standorten
  - Abbildung von 4 medizinischen Domänen an 9 Standorten
  - Bearbeitung von methodischen Themen an 7 Standorten
  - Administration, Koordination, Repositories an 2 leitenden Standorten



Systemaufbau an Annotations-Standorten

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

Dateihandling

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

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

Annotationsmethodik

- Aufbauend auf bestehenden Methoden
  - Annotationsguideline(s) – basierend auf bestehenden GL (AIDAVA, international) und Erfahrungen
  - 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 Annotion und De-Identifikation

Annotationseditor: INCEpTION

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

GeMTeX - Zusammenfassung

- Bereitstellung eines deutschen klinischen Text-Corpus
  - Umfangreich (60 k annotierte Dokumente)
  - Qualitätsgesichert
  - Offen
- Governance der MII
- 16 Partner
- Nachhaltigkeit durch Community-Integration