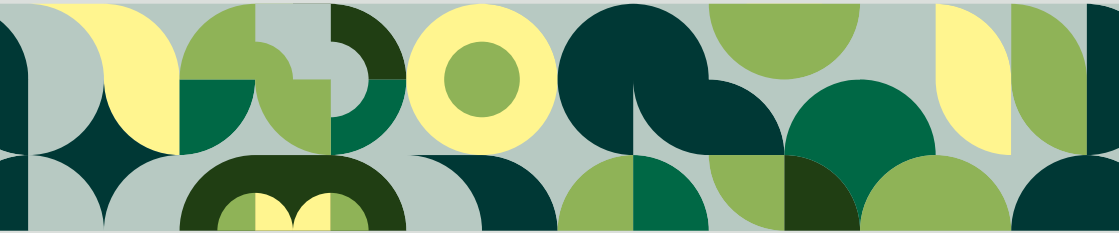
Pilotprojekt
»Neue Potenziale
für die digitale Lexikographie
des Deutschen«



Neue Potenziale für die digitale Lexikographie des Deutschen

„Neue Potenziale für die digitale Lexikographie des Deutschen“ (PDL) ist der Titel eines Pilotprojekts der Bayerischen Akademie der Wissenschaften. Während seiner zweijährigen Laufzeit sollen Grundlagen, Techniken und Strukturen für die Zusammenführung von lexikographischen Daten zur deutschen Sprache unter Berücksichtigung ihrer regionalen, historischen und sozialen Facetten erforscht werden.

Zentrales Ziel ist der Aufbau einer Forschungsdatenbank, um die weitgehend verstreuten und heterogenen Datenbestände systematisch zu integrieren und für fortgeschrittene linguistische Analysen und moderne algorithmische Verfahren verfügbar zu machen. Ebenso stellt die Entwicklung moderner digitaler Tools einen Schwerpunkt dar.

Das Ende 2025 gestartete Projekt wird in enger Kooperation mit den Wissenschaftsakademien der Akademienunion und in Zusammenarbeit mit dem Leibniz-Rechenzentrum der BAdW realisert. Ein wissenschaftlicher Beirat und die Expertise der Digital Humanities vor Ort unterstützen die Entwicklung.

Bereits in der Pilotphase wird eine umfangreiche Vernetzung mit weiteren Akademien, Universitäten und renommierten Forschungseinrichtungen angestrebt, um dem Charakter eines föderal organisierten Kompetenzzentrums zu entsprechen.

Fortschritte und Ergebnisse werden auf der Webseite des Projektes unter pdl.badw.de veröffentlicht.

Aufgaben, Themen und Forschungsbereiche des Pilotprojekts

Vernetzung
Förderung von Austausch und Kooperation der beteiligten Akademien

Forschungsdatenbank
Performante, verteilte, skalierbare und sichere Speicherung von Wörterbüchern und Korpora

Definition Datenformate
Standardisierte Verarbeitung von lexikographischen Daten des Deutschen auf Basis existierender Standards

Definition Austauschregeln
Erarbeitung klarer Regeln für die Datenweitergabe und -nutzung zur Stärkung der vernetzten Forschung

Aufbau Lexikographie-Portal
Bereitstellung von Daten und Tools für die wissenschaftliche Forschungsgemeinschaft

Wörterbuchkonsistenzprüfung
Algorithmen zur automatisierten Prüfung auf Fehler und heterogene redaktionelle Praktiken

KI-gestützte Workflows
Annotations-Copilots und Einsatz von Sprachmodellen zur Vereinfachung und Harmonisierung der Lemma-Erstellung

LLMs für Low-Resource-Sprachen
Transfer Learning und Modelltraining für Dialekte und Sprachvarietäten mit wenig Trainingsdaten

Vektor-/Embeddings-Tools
Systematische und standardisierte Berechnung von Embeddings und Speicherung in Vektordatenbanken

Visualisierungen/Wissensgraphen
Prägnante Kommunikation und Analyse von Wort- und Textzusammenhängen

Vernetzung

Aufgabenstellung

Eines der wichtigsten Ziele des Projekts: die Vernetzung der Akteure im Bereich der digitalen Lexikographie weiter zu stärken. Dazu gehören sowohl die Akademien, wie auch Universitäten, Forschungsinstitute und weitere Mitwirkende.

Mögliche Maßnahmen:

- Abgestimmte Kommunikationsstrategie
- Transparente Kommunikation der Projektziele und -fortschritte
- Veröffentlichung aller Tools (und Daten soweit möglich) zur Nachnutzung
- Lexikographie-Portal als Service-Angebot an die wissenschaftliche Gemeinschaft
- gemeinsame Veranstaltungen bzw. Workshops
- enge Zusammenarbeit mit der Akademienunion

Zeitplan

fortwährend

Q3/2026
Veranstaltungen

Stadtbibliothek Stuttgart
Tobias Fischer / Unsplash

Forschungsdatenbank

Zeitplan

12/25
Tests verschiedener DB-Systeme sowie APIs

03/26
Infrastruktur verfügbar

fortlaufend
Ergänzungen, Skalierung

Aufgabenstellung

In Zeiten knapper werdender Mittel für die Geisteswissenschaften besteht ein Lösungsansatz darin, Kompetenzen zu bündeln und Infrastruktur effizient gemeinsam zu nutzen.

Vor diesem Hintergrund werden gemeinsam mit dem Leibniz-Rechenzentrum Datenbank- und Speicherstrukturen erprobt bzw. realisiert, die in besonderer Weise dafür geeignet sind, lexikographische Daten und Korpora aufzunehmen.

Diese sollten zu einem späteren Zeitpunkt als Dienstleistung für andere wissenschaftliche Einrichtungen bereitgestellt werden, begleitet von technischer und inhaltlicher lexikographischer Expertise.

Zudem bietet die gemeinsame Datenbank einen deutschlandweit einmaligen Zugriff für analytische Forschungsfragen.

Definition Datenformate

Aufgabenstellung

Um die verteilten Wörterbuchdaten zentral speichern, auswerten und für weiterführende Tools verfügbar zu machen, wird ein einheitliches Datenformat benötigt.

Dessen Umfang und Gestalt wird maßgeblich von den Quellwörterbüchern bestimmt, da deren Tag-Strukturen bzw. Logiken abgebildet werden müssen.

Hierzu müssen komplexe und verlässliche Konvertierungsfunktionen implementiert werden. Gegenwärtig ist eine vollständige Kompatibilität mit dem Standard TEI Lex-0 vorgesehen.

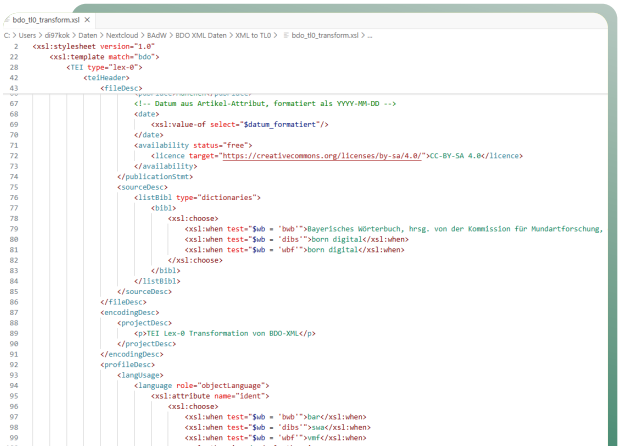
Für ein funktionales, akzeptiertes Format ist ebenfalls die Kooperation mit weiteren Arbeitsstellen essentiell.

Zeitplan

12/25
Transformation BDO XML
zu TEI Lex-0

06/26
Gemeinsames Format
der im Pilotprojekt
verarbeiteten Wörterbücher

09/27
OPTIONAL
Empfehlung für einen Format-
standard für Wörterbücher
des Deutschen



Code einer XSLT-Transformation zu TEI Lex-0
PDL

Definition Austauschregeln

Zeitplan
im Kontext mit Vernetzung

12/26
Entwuf eines
gemeinsamen Regelwerks

09/27
OPTIONAL
Finale Empfehlung

Aufgabenstellung

Die Arbeit an Wörterbüchern (und Korpora) hat eine jahrundertlange Tradition – und ebenso lange fallen bereits Daten an: handschriftliche Zettel, maschinengeschriebene Erhebungsbögen, Digitalisate, OCR-Erfassungen, und in modernen Formaten gespeicherte Daten. Erweitert wird der Daten-Reigen durch Algorithmen und Repositorien.

Doch trotz vielfältiger Kooperationen und gemeinsamer Projekte gibt es keine Standards für den Datenaustausch, abgesehen vor verschiedenen normierten Lizenzen.

Wie kann ein vertrauensvoller, qualitätswahrender, gemeinsamer Zugriff auf einen synergetischen Datenpool erfolgen, bei gleichzeitiger Wahrung der Einzelinteressen? Dieser Frage widmet sich dieser Themenschwerpunkt.

Aufbau Lexikographie-Portal

Aufgabenstellung

Um die im Rahmen des Pilotprojekts gesammelten Daten und realisierten Lösungen frei für andere zur Nutzung bereitstellen zu können, wird ein entprechendes Webportal aufgebaut. Es dient auch als Plattform für native Online-Tools.

Dieses Portal soll dabei nicht in Konkurrenz zu bestehenden lexikographischen Angeboten treten.

Es richtet sich hauptsächlich an die wissenschaftliche Fachcommunity und wird als offen gestaltete Ressource für neue Werkzeuge der digitalen Lexikographie konzipiert.

Zeitplan

03/26
Minimum Viable Product

12/26
Version mit erweiterten Funktionen online, inkl. erster Tools

fortlaufend
Ergänzungen, Skalierung

Zettelkasten
Jan Antonin Kolar / Unsplash

Wörterbuchkonsistenzprüfung

Zeitplan

12/26
Umsetzung von Prüfungsmechanismen am Beispiel „Bayerns Dialekte Online“

Aufgabenstellung

Wörterbücher waren und sind Langzeitprojekte. In der Folge ändern sich im Erfassungs- und Publikationszeitraum Personal, editoriale Standards, Eingabetools, Datenformate, Speicherformate, und viele weitere Rahmenbedingungen.

Eine automatisierte Konsistenzprüfung über bestehende, veröffentlichte Lemmata bzw. Bände kann Abweichungen vom Standard identifizieren und helfen, eine einheitliche Form sicherzustellen.

Verschiedene technische Ansätze sind denkbar, darunter sowohl regelbasierte Analysen wie auch der Einsatz von Sprachmodellen.

KI-gestützte Workflows

Aufgabenstellung

Dieses Modul beinhaltet zwei verschiedene Ansätze, wobei der erste vorrangig einen deterministischen Ansatz verfolgt, während der zweite sich die probabilistischen Fähigkeiten von Sprachmodellen zu Nutze macht.

Annotations-Copilot

Die Erfassung von lexikographischen Informationen für Wörterbücher erfordert Spezialwissen, hohe Präzision sowie große Kontinuität. Entsprechende Assistenzsysteme können helfen, die Bearbeitungszeit zu verkürzen und die Fehlerquote zu verringern, z.B. durch automatisierte Vorschläge, smarte Verweise oder Logikprüfungen.

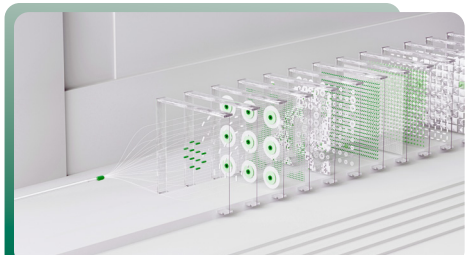
LLM-gestützte Lexikographie-Workflows

Schon heute lassen sich LLMs für die Erzeugung von Lexikoneinträgen nutzen - vermutlich mit gegenwärtig noch mehr Nach- als Vorteilen. Dennoch bieten LLM-gestütze Workflows großes Potential, die lexikographische Arbeit unter Aufrechterhaltung hoher wissenschaftlich-redaktioneller Standards effizienter und weniger fehleranfällig zu gestalten.

Beide Ansätze zeichnen sich jedoch dadurch aus, dass die bestehenden Workflows und verwendeten Tools in den Arbeitsstellen sehr vielgestaltig und individuell sind, und somit ein einmal erarbeiteter Workflow nicht ohne Weiteres übertragen werden kann.

Zeitplan

09/27
Umsetzung ausgewählter Tools am Beispiel „BDO“



Visualisierunge eines neuralen Netzwerks
Google Deep Mind / Unsplash

LLMs für Low-Resource-Sprachen

Zeitplan

Abhängig vom Kooperationsprojekt

Aufgabenstellung

Das Training von Sprachmodellen für Dialekte und andere Low-Resource-Sprachen sieht sich einem fundamentalen Problem gegenüber: Dem Mangel an geeigneten Trainingsdaten (d.h. umfangreiche schriftliche Quellen, räumlich und zeitlich möglichst kohärent).

Transfer Learning trainiert bestehende Modelle erneut, aber mit deutlich weniger Traningsdaten. Das ist vor allem dann erfolgsversprechend, wenn es große strukturelle Übereinstimmungen zwischen den ursprünglichen und den zusätzlichen Daten gibt, wie etwa im Beispiel Standardsprache-Dialekt.

Für dieses Modul wird eine externe Kooperation angestrebt.

Vektor- / Embeddings-Tools

Aufgabenstellung

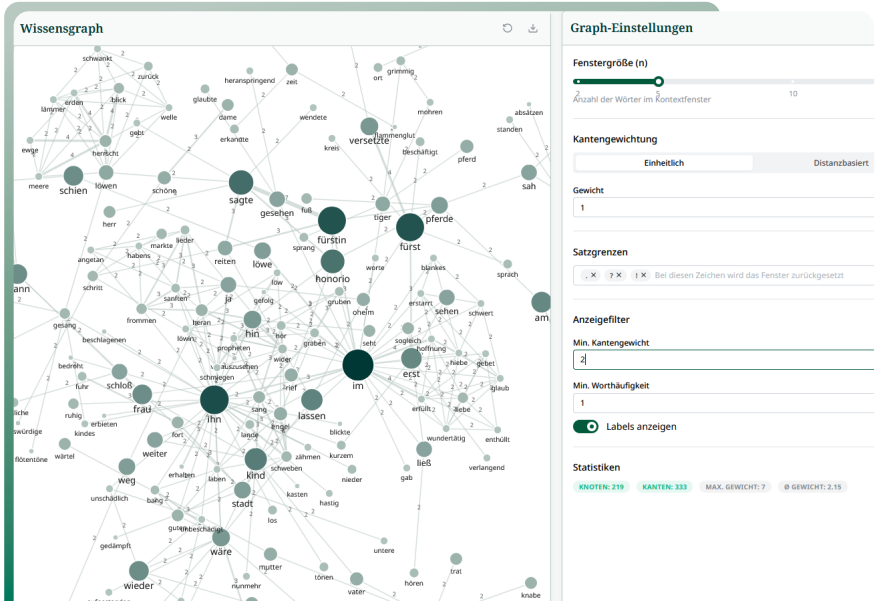
Statistisch-linguistische Analysen, die auf der Abbildung von Texten bzw. Korpora als hochdimensionale Vektoren basieren, müssen diese Vektor-Embeddings zunächst berechnen.

Dieser Prozess ist komputationell sehr aufwändig/teuer, und kann aufgrund dessen von kleinen Forschungsgruppen nicht geleistet werden. Daher werden solche Embeddings für ausgewählte Texte/ Korpora vorberechnet und zur Nutzung bereitgestellt.

In Zusammenarbeit mit anderen Forschungsstellen können auch komplexe Problemstellungen wie etwa historische Sprachstufen untersucht werden.

Zeitplan

09/26
Auswahl geeigneter Korpora
Berechnung beispielhafter Embeddings



Screenshot eines Stopword Removal/Graph Tools
PDL Lab

Visualisierungen/Wissensgraphen

Zeitplan

09/26
Bereitstellung im Lexikographie-Portal

Aufgabenstellung

Die existierenden Wörterbuchportale bieten bereits verschiedene Visualisierungen wie Wortwolken, Wortfeld-Netzwerke, und diachrone oder geographische Häufigkeitsverteilungen.

Können darüber hinaus noch weitere, innovative, informative Format entwickelt werden?

Als Objekte der Analyse kommen sowohl Wörterbücher wie auch Korpora in Frage. Ebenso soll die Möglichkeit geschaffen werden, Visualisierungen für von Nutzer*innen hochgeladene Texte zu generieren.

Ideen

Kontakte

Vernetzung

Sie interessieren sich für die Arbeit des Pilotprojekts?
Nehmen Sie gerne Kontakt mit uns auf:

Ansprechpartner

Wolfgang Huang
wolfgang.huang@adl.badw.de
+49 89 23031-1375

Dr. Fabian Simonjetz
fabian.simonjetz@adl.badw.de
+49 89 23031-1376

Webseite

pdl.badw.de

Neue Potenziale der digitalen Lexikographie des Deutschen

BAdW

BAYERISCHE AKADEMIE DER WISSENSCHAFTEN